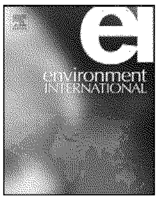
Environment International 92–93 (2016) 630–646

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint



Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review

Gbeminiyi O. Samuel[a], Sebastian Hoffmann[b], Robert A. Wright[c], Manoj Mathew Lalu[d], Grace Patlewicz[e,1], Richard A. Becker[f], George L. DeGeorge[g], Dean Fergusson[d], Thomas Hartung[a], R. Jeffrey Lewis[h], Martin L. Stephens[a,*]

[a] Johns Hopkins Center for Alternatives to Animal Testing, 615 N. Wolfe St., Baltimore, MD 21205, USA
[b] seh consulting + services, Stembergring 15, 33106 Paderborn, Germany
[c] William H. Welch Medical Library, Johns Hopkins University, 2024 E. Monument St., Suite 1-200, Baltimore, MD 21287, USA
[d] The Ottawa Hospital, The Ottawa Hospital Research Institute, Ottawa, Ontario K1Y 4E9, Canada
[e] DuPont Haskell Global Centers, 1090 Elkton Rd., Newark, DE 19711, USA
[f] Science and Research Division, American Chemistry Council, 700 2nd St., NE, Washington, DC 20002, USA
[g] MB Research Labs, 1765 Wentz Rd., Spinnerstown, PA 18968, USA
[h] ExxonMobil Biomedical Sciences, Inc., 1545 U.S. Highway 22 East, Room LA 350, Annandale, NJ 08801, USA

article info

Article history:
Received 6 August 2015
Received in revised form 9 March 2016
Accepted 9 March 2016
Available online 31 March 2016

Keywords:
Methodological quality
Reporting quality
Risk of bias
Scoping review
Toxicity studies

abstract

Assessments of methodological and reporting quality are critical to adequately judging the credibility of a study's conclusions and to gauging its potential reproducibility. To aid those seeking to assess the methodological or reporting quality of studies relevant to toxicology, we conducted a scoping review of the available guidance with respect to four types of studies: in vivo and in vitro, (quantitative) structure-activity relationships ([Q]SARs), physico-chemical, and human observational studies. Our aims were to identify the available guidance in this diverse literature, briefly summarize each document, and distill the common elements of these documents for each study type. In general, we found considerable guidance for in vivo and human studies, but only one paper addressed in vitro studies exclusively. The guidance for (Q)SAR studies and physico-chemical studies was scant but authoritative. There was substantial overlap across guidance documents in the proposed criteria for both methodological and reporting quality. Some guidance documents address toxicology research directly, whereas others address preclinical research generally or clinical research and therefore may not be fully applicable to the toxicology context without some translation. Another challenge is the degree to which assessments of methodological quality in toxicology should focus on risk of bias – as in clinical medicine and healthcare – or be broadened to include other quality measures, such as confirming the identity of test substances prior to exposure. Our review is intended primarily for those in toxicology and risk assessment seeking an entry point into the extensive and diverse literature on methodological and reporting quality applicable to their work.

1. Introduction

Research in toxicology, as in other fields, should be well-designed, rigorously conducted, and appropriately analyzed. These are key components of methodological quality. In clinical medicine, assessments of methodological or study quality typically focus on "risk of bias," i.e., the degree to which the design, conduct, and analysis of a study could potentially compromise confidence in its results by introducing systematic error in the magnitude or direction of the results (Higgins and Green, 2008). Risks of bias include, for example, failure to randomize study subjects to treatment groups or failure to "blind" outcome assessors to the treatment groups being assessed. Beyond risk of bias, methodological quality can also include other considerations. Within toxicology, these include adherence to standardized test guidelines and Good Laboratory Practices. Methodological quality is sometimes referred to as "reliability" in toxicology (Klimisch et al., 1997).

Striving for high standards of methodological quality should be coupled with similar rigor for reporting the results of research in the literature. Research should be reported accurately, thoroughly, and

[*] Corresponding author.
E-mail addresses: gsamuel4@jhu.edu (G.O. Samuel), Sebastian.hoffmann@seh-cs.com (S. Hoffmann), rwrigh32@jhmi.edu (R.A. Wright), manojlalu@gmail.com (M.M. Lalu), grace.y.tier@usa.dupont.com, Tier.Grace@epa.gov (G. Patlewicz), rick_becker@americanchemistry.com (R.A. Becker), degeorge@mbresearch.com (G.L. DeGeorge), dafergusson@ohri.ca (D. Fergusson), thartun1@jhu.edu (T. Hartung), r.jeffrey.lewis@exxonmobil.com (R.J. Lewis), msteph14@jhu.edu (M.L. Stephens).

[1] Present address: US EPA, National Center for Computational Toxicology (NCCT), 109 T. W. Alexander Dr., Research Triangle Park, NC 27711, USA.

http://dx.doi.org/10.1016/j.envint.2016.03.010
0160-4120/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

transparently. Reporting quality (sometimes referred to as "completeness of reporting") (Moher, 2015) is distinct from methodological quality but the two concepts overlap in a number of ways. Thorough reporting helps in the assessment of the methodological quality of a study. For instance, including only statistically significant results in a research paper is an example of both poor reporting and a risk of bias ("selective outcome reporting") (Guyatt et al., 2011a). Consequently, an appraisal of both methodological and reporting quality is essential to ensure that accurate information is derived from published research.

A considerable body of literature has addressed methodological and reporting quality, providing guidance not only on retrospectively assessing the quality of published studies but also on prospectively designing, conducting, analyzing, and reporting new studies. In this paper, we summarize both types of guidance. We address several types of studies of direct relevance to assessing the hazards and risks of environmental chemicals, namely, in vivo and in vitro (or mechanistic) studies, in silico studies (represented here by studies of (quantitative) structure–activity relationships ((Q)SARs), studies of physico-chemical properties, and observational human studies. In vivo studies examine effects on living animals, whereas in vitro studies examine effects on biomolecules, cells or tissues, from animals or humans. (Q)SARs are approaches that relate the properties of a chemical encoded in its molecular structure to a physical property or to a biological effect, e.g., toxicity. Studies of physico-chemical properties investigate, for example, a chemical's octanol–water partition coefficient, providing information that can guide subsequent toxicity testing. Human observational studies may explore the relation between human exposure to an environmental agent and a health effect. Such studies include various types (e.g., case–control, cohort, and cross-sectional).

For each study type, our aims were (1) to identify and summarize the available guidance on prospectively ensuring or retrospectively assessing methodological and reporting quality, and (2) to distill the common elements from this guidance. We adopted a scoping review approach. A scoping review "is a form of knowledge synthesis that addresses an exploratory research question aimed at mapping key concepts, types of evidence, and gaps in research related to a defined area or field by systematically searching, selecting, and synthesizing existing knowledge" (Colquhoun et al., 2014). Frameworks for the conduct of scoping reviews are emerging, and reporting guidelines are still in preparation (Colquhoun et al., 2014). Broadly speaking, scoping reviews identify the research topic; identify and select relevant studies; chart the data; collate, summarize, and report the results; and consult with relevant stakeholders (Arksey and O'Malley, 2005).

The literature on methodological and reporting quality has a rich history in clinical medicine and healthcare, thanks in part to an emphasis on evidence-based medicine. Our review emphasizes the relevance of this literature to toxicology and its diverse study types. It is intended primarily as an entry point into this literature for those in toxicology and risk assessment who wish to assess the methodological and reporting quality of research. Such assessments are usually retrospective (e.g., evaluating published studies) but can also be prospective (e.g., evaluating grant proposals). Apart from the assessment context, toxicologists have an obvious interest in ensuring the methodological and reporting quality of their own planned research.

Although toxicologists have grappled with issues of methodological and reporting quality over the years, some of the relevant terminology that has emerged primarily from other fields may be unfamiliar to toxicologists. Consequently, we provide a glossary of key terms in Table 1.

2. Methods

To retrieve published guidance on assessing or ensuring the quality of various types of studies relevant to toxicology, literature searches were devised and conducted with the aid of an information specialist (Appendix). Search strategies used a combination of controlled vocabulary and keywords adapted to each database searched. They were designed to achieve a balance of precision and recall in the results. There was no restriction on publication dates. Experts in toxicity research were consulted to identify any additional guidance.

Table 1
Glossary of key terms.

Allocation concealment: A process that it used to prevent selection bias. The person allocating subjects to experimental arms is unaware of which arm the subjects are being allocated until the moment of assignment. This prevents researchers from (unconsciously or otherwise) influencing the allocation of subjects (National Research Council, 2014; http://www.consort-statement.org/resources/glossary).

Attrition bias: Systematic differences in excluding study units between groups

Bias: Systematic deviation of the estimated intervention/exposure effect away from the "truth." This can be caused by inadequacies in the design, conduct, or analysis of an experiment, and produce deviations in either direction (i.e. under or over-estimate) (http://www.consort-statement.org/resources/glossary; handbook.cochrane.org/chapter_8/8_2_2_risk_of_bias_and_quality.htm).

Blinding (or masking): A set of procedures that keeps the participants and personnel involved in a study unaware of which intervention/exposure was received; this reduces the risk of performance bias. Similarly, outcome assessment can be blinded, so that personnel who assess outcome measures are unaware of the treatment allocation; this reduces the risk of detection bias (National Research Council, 2014).

Confounding bias: Systematic differences in factors potentially influencing the results between groups.

Detection bias: Systematic differences in the outcome assessment between groups

External validity: The extent to which a study provides a correct basis to generalize to other circumstances (Henderson et al., 2013).

Good Laboratory Practices (GLPs): A framework for study design, conduct, and oversight that reduces the risk of bias that can be associated with the adequacy of temperature, humidity, and other environmental conditions; experimental equipment and facilities; animal care; health status of animals; animal identification; separation from other test systems; and presence of contaminants in feed, soil, water, or bedding (National Research Council, 2014).

Internal validity: The extent to which the design and conduct of study minimizes bias and systematic error (Grimes and Schulz, 2002; http://www.consort-statement.org/resources/glossary).

Methodological quality: The extent to which the design and conduct of a study is likely to have prevented systematic errors (bias) (Olivo et al., 2008) and, as a result, identified "the truth" in its results and inferences. This term is quite similar to risk of bias.

Performance bias: Systematic differences introduced during the study.

Randomization: Randomly allocating an intervention under study across the comparison groups to ensure that group assignment cannot be predicted (National Research Council, 2014).

Reporting bias: Systematic omission of results in the study documentation/publication.

Reporting quality: Providing a complete and transparent description of the design, conduct, and analysis of a study (Moher et al., 1995). Also known as "reporting completeness."

Risk of bias: The risk of a systematic error or deviation from the truth in results or inferences. This term is interchangeable with internal validity (handbook.cochrane.org/chapter_8/8_2_2_risk_of_bias_and_quality.htm)

Scoping review: A form of knowledge synthesis that incorporates a range of study designs to comprehensively summarize and synthesize evidence with the aim of informing practice, programs, and policy and providing the direction for future research priorities (Colquhoun et al., 2014).

Selection bias: Systematic differences in the comparison groups.

Selective outcome reporting: The reporting of only selected results, not all results.

Searches with respect to in vivo and in vitro studies, (Q)SAR studies, and studies of physico-chemical properties were conducted on April 2, 3, and 4, 2015. PubMed and Embase were used to identify guidance for in vitro, in vivo, and (Q)SAR studies, whereas TOXLINE was used for physico-chemical studies. Separate searches were conducted in PubMed and the US Agency for Healthcare Research and Quality (AHRQ) reviews repository (http://www.effectivehealthcare.ahrq.gov) on April 5, 2015, to identify existing guidance on the quality of human studies. The EQUATOR Network website (http://www.equator-network.org/) was searched for guidance on reporting human observational studies. For human studies, the scope was limited to those studies that were observational rather than experimental, given the limited number of human experimental studies in the toxicological literature. Given the extensive literature on human observational studies, searches were targeted primarily towards guidance that had been summarized in reviews, rather than performing an exhaustive search of the primary literature.

For the literature searches, pre-determined criteria for inclusion of papers for each study type were used (Appendix). Publications were excluded, regardless of study type, if they covered content specific to a narrow sub-field (e.g. the methodological quality of animal research in critical care studies); were duplicates, editorials, or commentaries; emphasized topics other than guidance; were published in a language other than English; or were considered minor modifications of an approach published in an earlier document. We excluded references that focused on the methodological or reporting quality of systematic reviews, including meta-analyses, as these are appraised by different criteria than for individual studies (Liberati et al., 2009; Shea et al., 2007). Aside from human observational studies, reviews were excluded in favor of original sources. Additional details on the literature searches performed can be found in the Appendix.

Titles and abstracts were screened against the eligibility criteria by one author, who also accessed the full-text forms of promising papers to verify eligibility, searched the references of eligible publications to identify any additional pertinent papers, and extracted general characteristics (e.g., the objective) from each eligible paper. In addition, the proposed methodological and reporting criteria were extracted by two persons, who resolved any discrepancies through discussion.

Each eligible paper was categorized according to the type of toxicologically relevant study (e.g. human studies) and the type of quality (methodological or reporting) that it addressed. The identified documents vary in the extent to which they address methodological versus reporting quality. Papers that addressed both topics in a substantial manner were grouped into a "mixed guidance" category.

A number of methodological decisions were made in light of the ambiguities and inconsistencies in this rapidly evolving subject area. First, because terminology has not yet been standardized within and across disciplines (pre-clinical studies, toxicology, ecotoxicology), we grouped criteria that we considered sufficiently similar. For example, "independence of observations," "random outcome assessment" and "person assessing outcome has no knowledge of treatment assignment" were all considered to address the same aspect of detection bias, namely, blinding of outcome assessors. The criteria were described according to the most common description in the included studies.

Second, some criteria were considered by some authors as reporting elements and by other authors as methodological elements. We decided to avoid listing the same criterion in both categories, which we felt would be confusing. We resolved these situations by categorizing these criteria under methodological quality. Indeed, we considered all criteria proposed as essential elements of methodological quality to thereby qualify as essential elements of reporting quality, although we did not double-list these criteria under reporting quality.

We recognize that these methodological decisions may have introduced some subjectivity into the identification and categorization of criteria. Some level of subjectivity is unavoidable given the current state of the subject area. However, in view of our primary goal of providing an entry point into a rapidly evolving and consequently ambiguity-prone field, this was considered acceptable.

3. Results

The eligible papers derived from the search strategy are briefly summarized below under the type of toxicity study (in vivo and in vitro, (Q)SAR, physico-chemical, and human), together with the category of quality that they address (methodological, reporting, or mixed guidance).

Papers providing guidance aimed solely or primarily at toxicity studies are summarized first, followed by publications providing guidance aimed at other fields but which are nonetheless relevant to toxicology. Within this framework, papers are listed in chronological order. Where appropriate, each section concludes with a compilation of the most commonly proposed criteria for assessing the methodological and reporting quality for that type of toxicity study.

3.1. In vivo and in vitro studies

3.1.1. Literature search results

The literature search for guidance pertaining to in vivo and in vitro studies returned 3969 citations. Preliminary screening of the titles and abstracts of these citations yielded 82 papers for full-text review. Of these, 69 publications were excluded for a number of reasons (see Fig. 1).

Thirteen papers met the eligibility criteria. Five of these address methodological quality: Coecke et al. (2005), Hulzebos et al. (2010), Hooijmans et al. (2014), Rooney et al. (2014), and van Luijk et al. (2014). Two address reporting quality: Kilkenny et al. (2010) and Landis et al. (2012). Six provide mixed guidance: Klimisch et al. (1997), Festing and Altman (2002), Schneider et al. (2009), Hooijmans et al. (2010), Ågerstrand et al. (2010), and Beronius et al. (2014).

A review of the references in these 13 papers yielded six more papers that met the eligibility criteria and were included in this review. Four of these additional papers provide mixed guidance: Durda and Preziosi (2000), Küster et al. (2009), Macleod et al. (2009), and van

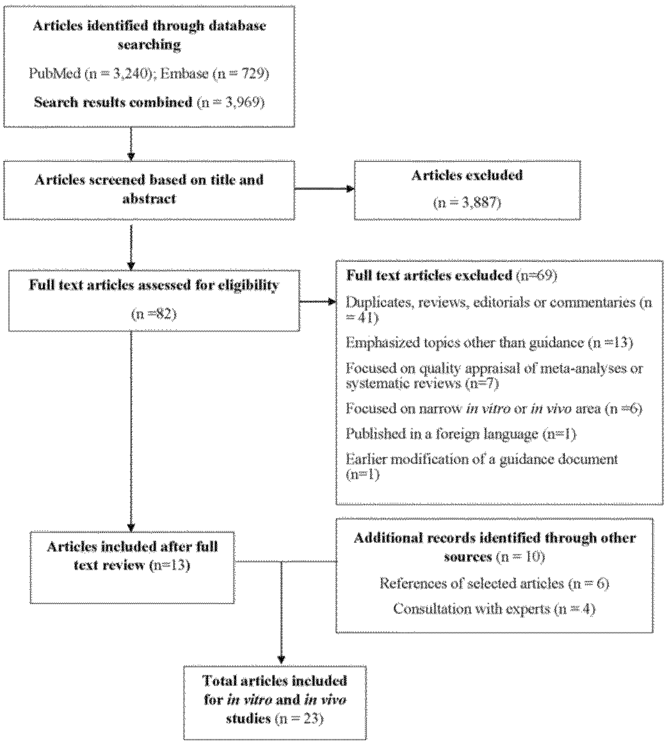

Articles identified through database searching
PubMed (n = 3,240); Embase (n = 729)
Search results combined (n = 3,969)
Articles screened based on title and abstract
Articles excluded
(n = 3,887)
Full text articles assessed for eligibility
(n =82)
Full text articles excluded (n=69)
Duplicates, reviews, editorials or commentaries (n =41)
Emphasized topics other than guidance (n =13)
Focused on quality appraisal of meta-analyses or systematic reviews (n=7)
Focused on narrow in vitro or in vivo area (n =6)
Published in a foreign language (n=1)
Earlier modification of a guidance document (n=1)
Articles included after full text review (n=13)
Additional records identified through other sources (n = 10)
References of selected articles (n = 6)
Consultation with experts (n = 4)
Total articles included for in vitro and in vivo studies (n = 23)

Fig. 1. Flow of included guidance documents on in vitro and in vivo studies.

der Worp et al. (2010). The remaining two address methodological quality: Hobbs et al. (2005) and Unger (2007). Another four publications were added based on consulting experts; these provide mixed guidance: OECD (1998), Code of Federal Regulations (2011), Maxim and van der Sluijs (2014), and National Research Council (2014). In total, 23 guidance documents for in vitro and in vivo studies were included, seven addressing methodological quality, two addressing reporting quality, and 14 addressing both methodological and reporting quality.

3.1.2. Methodological quality

Of the seven documents addressing the methodological quality of in vivo and/or in vitro studies, three were aimed primarily at the toxicological or environmental health communities and are listed first (Hobbs et al., 2005, Hulzebos et al., 2010, and Rooney et al., 2014). Four were aimed more broadly (Coecke et al., 2005, Unger, 2007, Hooijmans et al., 2014, and van Luijk et al., 2014).

1. Hobbs et al. (2005)

This paper presents the results of a rater experiment to improve the Australasian ecotoxicity database quality assessment scheme for aquatic toxicity data. The scheme consisted of 20 questions, which were applied independently by 23 ecotoxicologists to two research papers. As a result, refined criteria were proposed, potentially leading to more consistent ratings. These address study details such as duration of exposure, description of biological effect, use of appropriate controls, description of test acceptability criteria, and type of statistical model used. Each criterion is assigned a score and a cumulative score is calculated. Then an overall score is derived and used to characterize the data quality as "unacceptable" (b50%), "acceptable" (51 to 79%) or "high" (N80%).

2. Hulzebos et al. (2010)

Hulzebos et al. propose an Integrated Assessment Scheme (IAS) for evaluating the overall "adequacy" of (eco)toxicology data in meeting the information requirements under the European Union (EU) chemicals management system, the Regulation for Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). The IAS comprises three modules: (1) the "reliability" of the data, (2) the validity of the test method used, and (3) the regulatory need for the data. The validation principles of the Organization for Economic Cooperation and Development [see the (Q)SAR section on "Mixed guidance (methodological and reporting quality)"] were used to provide a harmonized set of criteria for assessing the three modules. Assessment categories identical to the Klimisch codes [see "Mixed guidance (methodological and reporting quality)"] are assigned to the evaluated information such that there are four possible categories in each of the three modules. The codes for reliability are R1 ("Reliable without restriction"), R2 ("Reliable with restriction"), R3 ("Non reliable"), and R4 ("Unassignable"). A similar rationale for classification applies to the remaining two modules, resulting in validity codes V1–V4 and regulatory need codes N1–N4. The various combinations of the three modules (e.g. R1–V2–N4) are assigned to three data adequacy conclusions: "adequate," "partly adequate," and "inadequate."

3. Rooney et al. (2014)

The National Toxicology Program Office of Health Assessment and Translation developed a seven-step framework for systematically reviewing environmental health questions to draw hazard identification conclusions. The seven steps are: formulate the problem and develop the protocol, search for and select the studies for inclusion, extract data from the studies, assess the quality or risk of bias of the individual studies, rate the confidence in the body of evidence, translate the confidence ratings into levels of evidence for a health effect, and integrate the evidence to develop hazard identification conclusions. The fourth step of the framework involves assessing the quality or risk of bias of individual studies. This step comprises seven risk of bias domains. These include selection bias (e.g. was exposure level adequately randomized?), confounding bias (e.g. did researchers adjust or control for other exposures that are anticipated to bias results?), performance bias (e.g. did researchers adhere to study protocol?), attrition/exclusion bias (e.g. were outcome data complete without attrition or exclusion from analysis?), detection bias (e.g. were the outcome assessors blinded to study group or exposure level?), selective reporting bias (e.g. were all measured outcomes reported?) and other (e.g. were statistical methods appropriate?). The overall risk of bias in the body of evidence is used as one of five properties that potentially influence the confidence in the body of evidence.

4. Coecke et al. (2005)

This paper proposes best practices in all aspects of the use of cells and tissues in vitro. The proposed Guidance on Good Cell Culture Practice (GCCP) provides standards for any work involving cell and tissue cultures, including the preparation of cells and tissues derived from humans and animals, characterization and maintenance of important characteristics, quality assurance, recording and reporting, safety, education and training, and ethics. The guidelines are applicable to in vitro testing used to satisfy regulatory requirements for chemicals.

5. Unger (2007)

This paper provides recommendations to improve the reliability and predictive capacity of preclinical translational research. According to the author, variability and bias are the principle challenges in designing, conducting and analyzing preclinical translational research studies. Recommendations to minimize variability include the use of a sample size large enough to overcome the variability in the model, derivation of disease-free animals of approximately the same age from a single source, and the identical care and handling of animals in all experimental groups. Recommendations to overcome biases in the design and analysis of translational research include: randomization and blinding, a prospective plan to manage missing data and outliers, use of rigorous statistical approaches, description of study limitations, and the substantiation of findings (i.e. to facilitate the independent reproduction of results in a subsequent study).

6. Hooijmans et al. (2014)

The Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE) developed a risk of bias tool for animal intervention studies. The tool is based primarily on the Cochrane Collaboration's risk of bias tool for randomized controlled trials. The resulting tool comprises 10 items, assessing six different types of bias: selection bias (e.g., was the animal allocation sequence adequately generated and applied?), performance bias (e.g. were the animals randomly housed during the experiment?), detection bias (e.g. were the animals selected at random for outcome assessment?), attrition bias (e.g. were incomplete outcome data adequately addressed?), reporting bias (e.g. are reports of the study free of selective outcome reporting?) and other biases (e.g. was the study apparently free of other problems that could result in high risk of bias?). Signaling questions were developed to assist quality evaluators in assigning a judgment of "low," "high" or "unclear" risk of bias to each item in the tool.

7. van Luijk et al. (2014)

In the context of improving the translation of animal data into clinical practice, the authors studied the risk of bias assessment in recent systematic reviews of preclinical animal studies as well as the actual risk of bias of the primary studies included in those reviews. Thirty-three systematic reviews and their associated primary studies were evaluated. The risk of bias assessment focused on the following four items (with the corresponding bias in parentheses): randomized study design (selection bias), blinding of investigator/caretaker (performance bias), blinding of outcome assessment (detection bias), and mentioning of drop-outs (attrition bias). The primary studies scored poorly (less than 25%) on each of these four elements, leading to the conclusion that the methodological quality of the primary animal studies should be improved.

3.1.3. Reporting quality

Two papers were identified that provide guidance on the reporting of in vivo and in vitro studies, both in the context of pre-clinical research.

1. Kilkenny et al. (2010)

The ARRIVE (Animals in Research: Reporting In Vivo Experiments) guidelines address the reporting of animal experiments. The guidelines were developed by researchers, statisticians, and journal editors, and funded by the United Kingdom-based National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs). The elements of the 20-item checklist are categorized under headings that follow the typical format of a scientific paper: Title, Abstract, Introduction, Methods, Results, and Discussion. The included items address ethical issues; study design; experimental procedures and specific characteristics of animals used; details of housing and husbandry; sample size; experimental, statistical, and analytical methods; and scientific implications, generalizability, and funding.

2. Landis et al. (2012)

These guidelines were proposed by major stakeholders in the US National Institute of Neurological Disorders and Stroke in order to improve the quality of reporting of animal studies. The authors reached consensus on a core set of reporting criteria that are recommended as prerequisites for authors of grant applications and scientific publications. The criteria comprise four items: randomization (e.g. data should be garnered and processed randomly), blinding (animal care-takers and investigators should be blinded), sample-size estimation (e.g. utilization of appropriate sample size), and data-handling (e.g. a priori description of inclusion and exclusion criteria). This guidance document served as the basis for the National Institutes of Health's core guidelines for reporting preclinical research (NIH, undated).

3.1.4. Mixed guidance (methodological and reporting quality)

With respect to in vivo and in vitro studies, 16 papers were identified that are a substantial mix of guidance on the methodological and reporting quality. Ten focus directly on the toxicological domain and are summarized first, while six provide guidance for pre-clinical studies.

1. Klimisch et al. (1997)

Klimisch et al. pioneered the quality assessment of toxicity and ecotoxicity studies in the context of the European Union's chemical regulation. Four reliability categories are proposed: “reliable without restriction,” “reliable with restrictions,” “not reliable,” and “not assignable.” Standard methods such as the OECD test guidelines are considered as reflecting the highest category, “reliable without restriction.” A mix of methodological and reporting criteria is presented for assessing non-standard studies. An example of methodological criteria includes description of the investigated outcomes. Reporting criteria include the specification of the test substance and information on dosing and/or data on animal feeding.

2. OECD (undated-a) and OECD (1998)

The OECD administers an influential test guidelines program that provides guidance on the design, conduct, analysis and reporting of in vivo and in vitro test methods. The individual test guidelines (OECD, undated-a) are internationally harmonized test methods for the evaluation of the safety of chemicals and chemical products. While not study protocols, the test guidelines include, inter alia, detailed recommendations and procedures for the selection of test species/strain, the assignment of unique identification numbers to each animal, and the determination of the frequency and endpoints for “in-life” observations.

In addition to test guidelines, the OECD (1998) has issued more general guidance in its “Principles of Good Laboratory Practice.” These principles address the performance of a study and the reporting of study results, as well as such diverse topics as the qualification of test facility personnel, quality assurance, and appropriate maintenance of laboratory apparatus. The Good Laboratory Practice (GLP) guidance on reporting addresses topics such as the test item and reference item, the sponsor and the test facility, the materials and test methods, and results (all information and data required by the study plan).

3. Durda and Preziosi (2000)

This paper proposes a two-step approach for assessing the quality of ecotoxicity studies. The first step entails assessing the compliance of studies with standardized toxicity testing and reporting protocols by applying various criteria that are organized into nine categories. These categories are hypothesis (e.g. endpoints appropriate for hypothesis); protocol (e.g. validate protocol, if not standardized); test compound (e.g. description of chemical species); dosing system (e.g. clearly described dose); test subjects (e.g. description of subject characteristics); controls (e.g. positive and/or negative controls); test environment (e.g. number of animals per test apparatus); statistical design (e.g. appropriate statistical model); and other considerations (e.g. results reproduced by others). In the second step a quality descriptor is assigned based on the degree of compliance with the established protocols. Five data quality descriptors are proposed, ranging from high (study carried out using standardized protocols) to not assignable (studies listed in short abstracts, secondary literature or otherwise lacking in documentation).

4. Schneider et al. (2009)

Schneider et al. propose the Toxicological Data Reliability Assessment Tool (ToxRTool) as a means of introducing more objectivity and consistency into the assignment of Klimisch categories to individual studies. The ToxRTool provides comprehensive criteria and guidance for these assignments. This software-based tool comprises two parts, one for in vivo studies and the other for in vitro studies. There are five evaluation criteria groups: (1) test substance identification, (2) test system characterization, (3) study design description, (4) study results documentation, and (5) plausibility of study design and data. Studies are assigned scores that are translated into Klimisch categories. Criteria that are considered essential (e.g. test substance identification and test concentration description) are given greater weight in the evaluation. The ToxRTool is nested within a Microsoft Office Excel® 2003 file that contains spreadsheets for the reliability evaluation of in vivo and in vitro toxicity studies, optional documentation of observations with importance to relevance (e.g. was the study conducted according to recent OECD or EU guidelines?), as well as detailed explanations of the criteria. The tool is available for download at https://eurl-ecvam.jrc.ec.europa.eu/about-ecvam/archive-publications/toxrtool

5. Küster et al. (2009)

This paper proposes quality criteria for literature data used in the environmental risk assessment of pharmaceuticals (human and veterinary) to increase clarity in this risk assessment process. The risk assessment involves appraising the submitted literature data for completeness of reporting and plausibility, as well as adherence to current fate and ecotoxicological standards. Documentation requirements are presented for various study types. The quality of data can be classified into one of three categories: (1) Data that are reliable without restriction according to the European Medicine Evaluation Agency (EMEA) guideline (studies carried out according to internationally accepted test guidelines [e.g. OECD]). (2) Data are reliable with restriction according to the EMEA guideline (e.g. studies in which test parameters documented are not compliant with the corresponding test guideline, but are sufficient to evaluate the data. (3) Data are not reliable according to instructions in the EMEA guideline (e.g. insufficiently documented studies).

6. Ågerstrand et al. (2011)

The authors developed a set of evaluation and reporting criteria to improve the scientific basis of environmental risk assessments for pharmaceuticals. A two-dimensional evaluation is proposed addressing both relevance and reliability. Twelve criteria are proposed to determine low or high relevance, while reliability is rated as low or high using 10 categories, such as "purpose and endpoint" or "test organism," by applying 63 individual criteria. Combining both ratings, data can be assigned to one of four fields, which determine the weight the data should receive in risk assessment. For example, data with both high reliability and relevance should have a high weight in risk assessment.

7. EPA (undated-a) and Code of Federal Regulations (2011)

The US Environmental Protection Agency (EPA) has issued standardized test guidelines that are intended to encourage the performance of high quality studies that are both relevant and reliable for determining potential hazards and dose response for regulatory evaluations (EPA, undated-a). They are supplemented with the agency's own GLP guidance (Code of Federal Regulations, 2011). The GLP guidance specifies standards intended to ensure the quality and integrity of in vivo and in vitro data submitted to the agency in support of regulatory evaluations for pesticide products and chemicals. These standards cover such diverse topics as test facility organization and personnel (training and responsibilities), quality assurance, facilities, performance of laboratory equipment and instruments, justification of the test method, design and performance of the study, and reporting of study results. Responsibilities of the quality assurance program include maintenance of approved study plans and standard operating procedures, verification of study plans to ensure compliance with GLP principles, and inspection of facility and process in accordance with GLP principles. Thus, the EPA's guidelines and GLP standards mirror those of the OECD in addressing, explicitly or implicitly, important elements of methodological and reporting quality.

8. Maxim and van der Sluijs (2014)

This paper proposes the "Qualichem in vivo" tool for evaluating the quality of in vivo studies used in chemical health risk assessments. The tool parses quality appraisal into four domains: technical (e.g. technical errors resulting from imprecise tools or measurement methods), methodological (e.g. the use of best available scientific knowledge and practices during the research protocol), normative (e.g. the interpretation of raw data and conclusions about level of evidence), and communicational (e.g. comprehensive reporting of research). Forty-five quality criteria were developed and divided into two general categories: "Protocol" and "Results." The "Protocol" section addresses technical and methodological issues (e.g. check of substance properties, check of storage conditions, handling of experimental animals, and precision of effects measurements). The "Results" section also addresses technical and methodological issues (e.g. statistical methods used, status of peer review, and coherence with literature) as well as issues related to communicational quality (e.g. result reporting) and normative quality (e.g. causal interpretations and interpretations based on existing scientific knowledge). The tool was evaluated using two case studies involving Bisphenol A.

9. Beronius et al. (2014)

Beronius et al. propose criteria for assessing reliability and relevance of non-standard in vivo studies. A two-tiered approach for assessing reliability was developed. The 11 Tier I reliability criteria address, for example, appropriate substance identity description and information on the animals used, such as the species, sex and age. The reliability of studies that satisfy all of the Tier I criteria are then evaluated in more detail in Tier II, which is available as a web-tool. The proposed 32 Tier II reliability criteria are grouped in seven categories, such as "purpose" and "test compound." Finally, relevance is evaluated, using eight items that comprise aspects such as the relevance of the route of administration for human exposure and the appropriateness of exposure timing for the investigated endpoints. Furthermore, the authors propose a reporting checklist with items important for the evaluation of reliability and relevance for risk assessment purposes.

10. National Research Council (2014)

This National Research Council (NRC) report provides an overview of general issues associated with the EPA Integrated Risk Information System (IRIS) assessments. The report addresses evidence identification and integration for hazard evaluation. Chapter 5 focuses on a critical part of the systematic review process: the assessment of individual studies that are selected for inclusion in a review. The best practices for evaluating clinical and epidemiologic studies, animal toxicology studies, and mechanistic studies in the systematic review process are discussed. The authoring committee emphasizes the need for EPA to assess the "risk of bias" in individual studies. The report identifies the various types and sources of bias within a study, including lack of randomization, blinding, inclusion and exclusion criteria, statistical power, outcome assessment, use of clinically relevant animals, and inconsistent standards for reporting. It is highlighted that in order to overcome these pitfalls, these quality criteria should be included in GLPs that apply to animal studies. Acknowledging that only a few tools are available for risk of bias evaluation of mechanistic toxicity studies, several approaches to overcome this lack are considered. Recommendations proposed by the report include, inter alia, (1) that the EPA should advance the tools for assessing the risk of bias in different types of studies (human, animal and mechanistic) used in IRIS assessments and develop tools for assessing risk of bias for in vitro studies; (2) that the EPA should select a method for the evaluation of individual studies that is transparent, reproducible, and scientifically defensible; and (3) that a coordinated effort of many stakeholders is needed to improve study reporting.

11. Festing and Altman (2002)

Festing and Altman propose a guideline to support investigators using animals with a focus on experimental design and statistical data analysis. Among other aspects, they highlight the importance of randomization, blinding, sample size calculation and appropriate statistical analysis. In addition, guidance is given on the presentation of results and on what information about animals and their environment should be reported.

12. Macleod et al. (2009)

This paper sets out a series of standards to reduce bias in the design, conduct and reporting of animal experiments modeling human stroke. The authors advocate the general adoption of experimental standards to ensure decision-making is based on high quality, unbiased data and further advocate that these standards be described in the "methods" sections of scientific publications. In total, eight standards were proposed: (1) the species, strain/sub-strain, and source of the animals used; (2) the sample size calculation; (3) inclusion and exclusion criteria; (4) the method of randomization; (5) allocation concealment; (6) reporting of animals excluded from analysis (including rationale); (7) blinded assessment of outcome; and (8) reporting of potential conflict of interest.

13. van der Worp et al. (2010)

This paper investigates the inadequacies of preclinical studies with regard to internal validity, external validity, and publication bias in favor of positive studies. The objective was to provide practical strategies to improve failed translational animal research. Four types of bias threatening internal validity (see Table 1) were defined and solutions were proposed to address selection bias (randomization and allocation concealment), performance bias (blinding), detection bias (blinding), and attrition bias (blinding and intention-to-treat analysis). In addition six common causes for reduced external validity (see Table 1) were outlined, e.g., the use of young and healthy animals for elderly disease, the use of models with insufficient similarity to the human condition, and the use of toxic or not-tolerated doses. To prevent publication bias, aspects of study quality to be reported in manuscripts were proposed, including: the sample size calculation, eligibility criteria, treatment allocation to experimental groups, allocation concealment, blinding, flow of animals, control of physiological variables, control of study conduct, and statistical methods.

14. Hooijmans et al. (2010)

The Gold Standard Publication Checklist provides detailed guidelines on the proper reporting (and design) of animal experiments. It is intended to improve the quality of research involving animals, to help researchers to replicate results, to reduce the number of animals used in research, and to improve animal welfare. The checklist comprises several items under four categories similar to those of the ARRIVE guidelines (see above): Introduction, Methods, Results and Discussion. The guidelines recommend that the methods section addresses the following topics: the experimental design used; the experimental groups and controls used (such as species, genetic background, housing and housing conditions, and nutrition); the ethical and regulatory principles followed; the intervention employed (such as dose and/or frequency of intervention, and administration route); and the desired outcome (such as descriptions of parameters of interest and statistical methods).

3.1.5. Criteria summary

There is a substantial literature on the methodological and reporting quality of in vivo and in vitro studies, although only one paper (Coecke et al., 2005) addresses in vitro studies exclusively. Much of this literature is focused directly on toxicity studies. Moreover, the guidance aimed more generally at preclinical studies has clear relevance to toxicity studies, with appropriate translation.

Fifteen criteria for addressing the methodological quality of in vivo and in vitro studies were proposed in at least four (~20%) of the 19-relevant documents and these are listed in Table 2. Eight of these 15 criteria are readily aligned with standard risk of bias categories and are grouped under those headings in the table (selection bias, performance bias, detection bias, attrition bias, reporting bias, and confounding bias) for convenience. However, seven criteria did not fit this framework (e.g., optimal time window used). Eight items were proposed in fully 10 (~50%) or more of the 19 documents surveyed. These eight criteria are a mix of risk of bias concerns (randomized allocation of subjects, blinding of researchers and outcome assessors, complete outcome data, and selective outcome reporting) and other criteria (information on the test organism/system, test substance/treatment details, and appropriate/controlled exposure).

Four criteria were proposed in less than 20% of the relevant documents (and are not listed in Table 2), comprising the definition of inclusion/exclusion criteria, the requirement of random outcome assessment, the requirement of identical experimental conditions during the study (pertinent to minimization of performance bias) and the control of biasing co-exposures.

With respect to reporting quality, there was substantial consistency across the guidance documents (Table 3). Eight of the 12 total criteria were proposed in at least 50% of the 12 documents surveyed. The three criteria proposed in more than 80% of the guidelines are a study design description; information on housing, feeding, and maintenance conditions; and a justification and description of statistical methods. The least frequent criteria were the requirement of describing the scientific background and the inclusion of an ethical statement (both in 27% of the documents).

It bears repeating (see Methods section) that we consider all criteria proposed as essential elements of methodological quality to thereby qualify as essential elements of reporting quality, although we did not double-list these criteria under reporting quality.

Further information on the subject of methodological or reporting quality, at least with respect to in vivo studies, can be found in Henderson et al. (2013), Krauth et al. (2013), Bailoo et al. (2014), and O'Connor and Sargeant (2014), papers that were identified in our literature search but were excluded from eligibility because they were reviews of the subject, not primary guidance documents.

3.2. (Q)SAR studies

3.2.1. Literature search results

Our literature search for guidance on the methodological and reporting quality of (Q)SAR studies returned 5990 citations. Preliminary screening of titles and abstracts yielded 44 papers for full-text review (see Fig. 2). Forty-three of these were excluded because they did not provide relevant guidance. Only one publication (Hulzebos et al., 2010) met the eligibility criteria. This paper is summarized in the section on In vivo and in vitro studies, to which it is also relevant. However, a review of the reference section of Hulzebos et al. (2010) yielded two other relevant documents, the OECD guidance document on the validation of (Q)SAR models (OECD, 2007) and the European Chemicals Agency (ECHA) guidance document for (Q)SAR studies (ECHA, 2008). These two papers are summarized below, both in the "mixed guidance" category.

3.2.2. Mixed guidance (methodological and reporting quality)

1. OECD (2007)

A 2002 workshop hosted by the European Centre for Ecotoxicology and Toxicology of Chemicals and organized by the International Council of Chemical Associations and the European Chemical Industry Council was held in Setubal, Portugal. It brought together a diverse group of international stakeholders to develop proposals for guidance and criteria for the regulatory acceptance of (Q)SARs. Six guiding principles for the development and application of (Q)SARs for regulatory purposes were proposed, which became known as the Setubal Principles (Jaworska et al., 2003). These were subsequently discussed and endorsed by the OECD and are now known as the OECD Principles for (Q)SAR Validation (OECD, 2007). These principles provide a framework for determining the scientific validity of (Q)SAR models for regulatory purposes. The five principles are: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; and (5) a mechanistic interpretation (if possible). Preliminary guidance to interpret these principles was developed by the European Commission's Joint Research Centre (Worth et al., 2005) and subsequently incorporated into OECD guidance (OECD, 2007).

Reporting formats to capture (Q)SAR information were developed under the auspices of the then EU Technical Committee for New and Existing Substances QSAR Working Group. Two reporting formats in particular are worth noting — the (Q)SAR Model Reporting Format (QMRF) and the (Q)SAR Prediction Reporting Format (QPRF) (OECD, 2007; ECHA, 2008). The QMRF contains information on the source, type, development, validation, and possible applications of the model. These types of information are reflected in Table 4. The QPRF describes the evaluation of a specific substance by a specific (Q)SAR model described in the associated QMRF. It addresses the evaluation of the reliability of the prediction. The type of information captured in the QPRF includes a description of how well the substance falls within the defined domain of applicability and the extent to which there is agreement between the (Q)SAR predictions and the experimental data for relevant analogues.

Table 2
Commonly proposed criteria for assessing the methodological quality of in vivo and in vitro studies from the guidance documents summarized in the “Methodological quality” and “Mixed guidance (methodological and reporting quality)” subsections of the “In vivo and in vitro studies” section. These guidelines may propose a wider array of quality criteria; here, we list only those that are most commonly proposed.

Guidance	Selection bias			Performance bias	Detection bias	Attrition bias	Reporting bias	Confounding bias	Appopriate statistical methods		Appropriate/ controlled exposure (incl. characterization)	Optimal time window used	Statement of conflict of interest/ funding source	Test substance/ treatment details	Test organism/ system
	Baseline characteristics similarity/ appropriate control group selection	Allocation concealment	Randomization	Blinding of researchers	Blinding of outcome assessors	Complete outcome data	Selective outcome reporting	Account for confounding variables	Sample size determination	Statistical analysis					
Beronius et al. (2014)[a]	✓	–	✓	–	–	✓	✓	–	✓	✓	✓	✓	✓	✓	✓
Hooijmans et al. (2014) (SRYCLE)[a]	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	✓	✓	–	✓
Maxim and van der Sluijs (2014) (Qualichem In Vivo)[a]	–	–	✓	✓	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	✓
National Research Council (2014)[c]	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	–	✓	✓	✓
Rooney et al. (2014) (NTP/OHAT)[c]	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	–	–
van Luijk et al. (2014)[a]	–	–	✓	✓	✓	✓	–	–	–	–	–	–	–	–	–
Ågerstrand et al. (2011)[c]	✓	–	✓	✓	✓	✓	–	✓	–	✓	✓	✓	–	✓	✓
Hooijmans et al. (2010) (GSPC)[a]	✓	–	✓	✓	✓	✓	–	–	✓	–	–	–	–	–	✓
Hulzebos et al. (2010) (IAS)[c]	–	–	–	–	–	–	–	–	–	–	✓	–	–	✓	✓
van der Worp et al. (2010)[a]	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–
Küster et al. (2009)[c]	–	–	–	–	–	✓	✓	–	–	–	✓	✓	–	✓	✓
Macleod et al. (2009)[a]	–	✓	✓	✓	✓	✓	–	–	✓	–	–	–	✓	–	✓
Schneider et al. (2009) (ToxRTool)[c]	–	–	✓	–	–	–	✓	–	–	–	✓	–	–	✓	✓
Unger (2007)[a]	✓	✓	✓	✓	✓	✓	–	✓	–	✓	–	–	–	–	✓
Coecke et al. (2005) (GCCP)[b]	–	–	–	–	–	–	✓	–	–	–	–	✓	–	–	✓
Hobbs et al. (2005)[c]	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓
Festing and Altman (2002)	✓	–	✓	✓	✓	✓	–	✓	✓	✓	–	–	–	–	✓
Durda and Preziosi (2000)[c]	✓	–	✓	✓	✓	–	–	–	✓	✓	✓	–	–	✓	✓
OECD (undated-a,1998), EPA (undated-a), CFR (2011)[c]	n.a.														
Klimisch et al. (1997) (Klimisch System)[c]	–	–	–	–	–	–	✓	✓	–	–	✓	–	–	✓	✓
% of total (N = 19)	47	32	74	63	63	63	53	47	42	47	58	29	26	53	84

Key—a: Guideline applies to in vivo studies only; b: Guideline applies to in vitro studies only; c: Guideline applies to both in vivo and in vitro studies; n.a.: not applicable. Supplemental materials reviewed for guideline appraisal in Schneider et al. (2009) and Maxim and van der Sluijs (2014).

Category descriptions (see also glossary in Table 1)—Account for confounding variables: This is very context depending. In an animal study of endocrine disruption, bedding material potentially containing phytoestrogens should be the same for all groups. Appropriate/controlled exposure (including characterization: It needs to be ensured that all subjects are treated/exposed in the same way, e.g., by controlling the food consumption per animal in a feeding study. Appropriate statistical methods: Appropriateness of statistical methods of experimental design and data analysis has to be demonstrated/justified. Baseline characteristics similarity/appropriate control group selection: Control and treated groups are similar at the start of the study, e.g. sex ratio, weight and age distribution. Complete outcome data: Accounting for all included study units. Optimal time window used: This refers to the age and status (e.g., pregnancy or disease status) of the animals. In a developmental toxicity study, for example, the exposure should take place during the most appropriate gestation days. In cell culture experiments, the cells should be exposed at their optimal developmental state, e.g., a t c o fluency, or within certain cell passage numbers, for which the stability of the karyotype is guaranteed. Statement of conflict of interest/funding source: C o flicts or funding by bodies with vested interests may result in (un-)conscious biases during the entire study, from planning to publication. Test organism/ system: The animal type/strain or the cell system needs to be stated, e.g. using different cell batches may introduce bias. Test substance/treatment details: The test substance identity should be known, including possibly interfering impurities. Treatment details should be known, in order to assess issues such as optimal time window used.

Table 3
Commonly proposed criteria for assessing the reporting quality of in vivo and in vitro studies from the "Reporting quality" and "Mixed guidance (methodological and reporting quality)" subsections of the "In vivo and in vitro studies" section. These guidelines may assess a wider array of criteria; here we list only the most commonly proposed quality criteria.

Guidance	Description of scientific background	Description of study purpose/ objective	Justification for model	Study design description	Defined experimental outcomes	Information on housing and feeding/ maintenance conditions	Ethical statement	Statistical analysis	Description of measurement precision and variability	Results description	Dose/concentration-response consideration	Result interpretation/ discussion
Beronius et al. (2014)[a]	–	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	✓
Maxim and van der Sluijs (2014) (Qualichem In Vivo)[a]	–	–	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
National Research Council (2014)[c]	No specific reporting criteria, but the importance of complete and accurate reporting is acknowledged											
Landis et al. (2012)[a]	–	–	–	–	✓	–	–	✓	–	✓	–	–
Ågerstrand et al. (2011)[c]	–	✓	✓	✓	–	✓	–	✓	–	✓	✓	✓
Hooijmans et al. (2010) (GSPC)[a]	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	–	✓
van der Worp et al. (2010)[a]	✓	–	–	–	–	–	–	✓	–	–	–	–
Kilkenny et al. (2010) (ARRIVE)[a]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓
Macleod et al. (2009)[a]	All requirements relate to methodological quality and are presented in Table 2											
Schneider et al. (2009) (ToxRTool)[c]	–	✓	✓	✓	✓	✓	–	✓	✓	✓	–	–
Küster et al. (2009)[c]	–	–	–	✓	✓	✓	–	✓	–	–	–	–
Festing and Altman (2002)[a]	–	✓	✓	✓	–	✓	–	✓	✓	✓	–	✓
Durda and Preziosi (2000)[c]	–	✓	–	✓	–	✓	–	✓	–	–	✓	–
OECD (undated-a, 1998), EPA (undated-a), CFR (2011)[c]	n.a.											
Klimisch et al. (1997) (Klimisch system)[c]	–	–	✓	✓	✓	✓	–	✓	–	✓	✓	–
% of total (N = 12)	25	58	58	83	67	83	25	100	42	75	42	50

Key —a: Guideline applies to in vivo studies only; b: Guideline applies to in vitro studies only; c: Guideline applies to both in vivo and in vitro studies; n.a.: not applicable.

Although this OECD guidance was developed to support efforts of (Q)SAR model application to regulatory purposes, it is also relevant to the context of assessing the methodological and reporting quality of individual (Q)SAR studies.

2. ECHA (2008)

ECHA administers the REACH regulation. ECHA's (Q)SAR technical guidance is aligned with the aforementioned OECD guidance. Under REACH, the QMRF and QPRF are used to ensure transparency (unambiguous reports of estimation methods, prediction, and reasoning); consistency; and acceptability (report of all relevant information to assess the adequacy and completeness of (Q)SAR information for a given substance or endpoint). These formats are also used to satisfy the need for classification, labeling, and risk assessment.

3.2.3. Criteria summary

Guidance on the methodological and reporting quality of (Q)SAR studies is limited but authoritative, coming mostly from the OECD and the ECHA (Table 4). Given that this guidance was developed to support efforts at model validation, some components may need to be appropriately translated to the context of assessing the quality of individual (Q)SAR studies employing a given validated model. Indeed, some components may not even apply outside of the validation context. For example, validation principle four consists of statistical validations and relates to issues such as goodness of fit, sensitivity, internal validation techniques, and training and test sets. Considerable evidence on these issues would need to be marshaled in the context of an actual validation exercise, but such data could simply be referenced in the context of assessing the methodological and reporting quality of an individual application of a given model.

It remains to be determined how these (Q)SAR quality elements translate to the risk of bias framework from clinical medicine. One of the components of a defined (Q)SAR endpoint is data quality and variability (Table 4). How similar is the assessment of data quality and variability in this context compared to the assessment of risk of bias in clinical medicine? To what extent are (Q)SAR developers assessing the methodological and reporting quality of the underlying experiments on which their models are based, to avoid the familiar problem of "garbage in, garbage out"?

3.3. Studies of physico-chemical properties

3.3.1. Literature search results

Ninety-four citations were returned in the literature search for guidance on the methodological and reporting quality of studies of physico-chemical properties (Fig. 3). Preliminary screening of titles and abstracts yielded two documents for a full-text review: EPA (undated-b) and Arts et al. (2008). The former met the eligibility criteria while the latter was excluded following the full-text review, as it was not a primary guidance document. However, Arts et al. (2008) made reference to OECD guidance on test methods used in the identification and characterization of hazards from chemical substances. Based on this observation, the OECD website was successfully searched to capture guidance on physico-chemical properties of chemicals. EU guidelines similar to those of the EPA were identified through consulting experts. All three sets of guidance fell into the mixed category.

3.3.2. Mixed guidance (methodological and reporting quality)

1. OECD (undated-b)

The OECD provides detailed guidance for conducting and reporting physico-chemical test methods. About 25 different test guidelines describe procedures to determine approximately 20 physico-chemical properties, including water solubility, dissociation constants in water, and hydrolysis as a function of pH. The guidelines also outline the types of information that should be reported. For example, the test

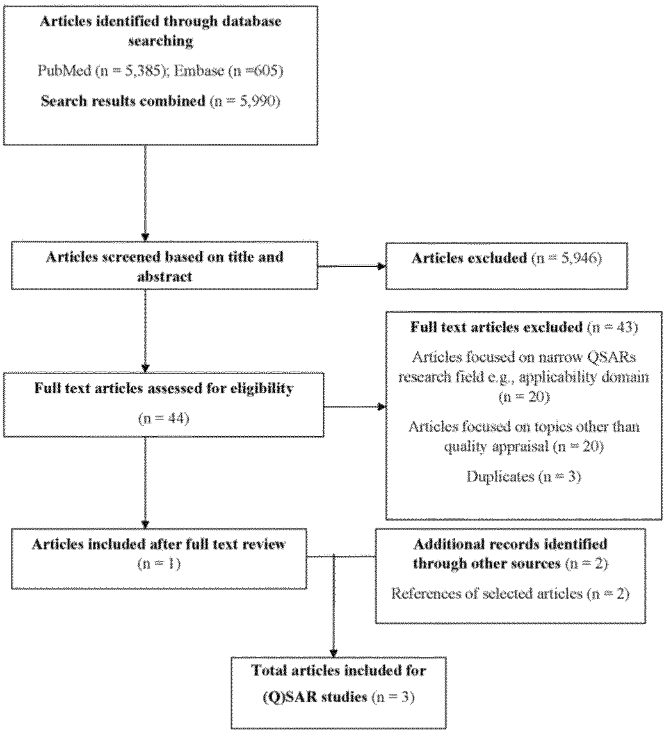

Articles identified through database searching
PubMed (n = 5,385); Embase (n =605)
Search results combined (n = 5,990)
Articles screened based on title and abstract
Articles excluded (n = 5,946)
Full text articles excluded (n = 43)
Articles focused on narrow QSARs research field e.g., applicability domain (n = 20)
Articles focused on topics other than quality appraisal (n = 20)
Duplicates (n = 3)
Full text articles assessed for eligibility
(n = 44)
Articles included after full text review (n = 1)
Additional records identified through other sources (n = 2)
References of selected articles (n = 2)
Total articles included for (Q)SAR studies (n = 3)

Fig. 2. Flow of included guidance documents on (Q)SAR studies.

guideline on vapor pressure describes eight different measuring methods that can be applied in different vapor pressure ranges. For each of these eight methods, the guideline provides details on the required apparatus, step-by-step technique, data collection, and data reporting.

2. European Union (2008)

The European regulation No. 440/2008 includes test methods for a wide range of physico-chemical properties. The majority of the methods mirror the respective OECD test guidelines.

3. EPA (undated-b)

The US EPA issues guidelines for testing pesticides and toxic substances and developing test data for submission to the Agency for review under various statutes. Guidance relevant to assessing the methodological and reporting quality of physico-chemical properties is captured under final test guidelines 830.6302 through 830.7950.

3.3.3. Criteria summary

As with the guidance on (Q)SAR studies, guidance on the methodological and reporting quality of studies of physico-chemical properties is limited but authoritative, in this case coming from the OECD, EU, and the US EPA. The guidance comprises a set of test procedures intended primarily for the prospective design, conduct, analysis, and reporting of physico-chemical studies. However, this guidance can also be applied to the retrospective appraisal of such studies. It is endpoint-specific, apparently with no explicit standards applicable to assessing the methodological and reporting quality of all such studies. However, further guidance can be obtained from the OECD, EU, and EPA GLP principles, which, though discussed above in the context of in vivo and in vitro studies, should also be instructive for other types of studies.

It has yet to be determined to what extent the clinical risk of bias framework translates to the domain of physico-chemical studies in toxicology.

Table 4
Summary of the OECD principles and their components for guiding (Q)SAR validation (OECD, 2007), which can also be applied (appropriately translated) to assessing the methodological and reporting quality of individual (Q)SAR studies. Similar guidance is available from ECHA (2008).

	Validation principle	Components
1.	A defined endpoint	Relevant experimental factors (e.g. species) Endpoint: environmental fate (e.g. biodegradation) Ecological effects (e.g. acute fish toxicity) Human health effects (e.g. acute oral toxicity) Physico-chemical properties (e.g. melting point) Dependent variable Endpoint units Experimental protocol Data quality and variability
2.	An unambiguous algorithm	Type of model Explicit algorithm Descriptors, descriptor selection Algorithm and descriptor generation Software for algorithm and descriptor generation Chemical/descriptor ratio
3.	A defined domain of applicability	Structural fragment domain Descriptor domain Mechanistic domain (mode of action, range of activity) Metabolic domain (transformation or metabolism)
4.	Statistical validations	Goodness of fit Accuracy Sensitivity Specificity Internal validation techniques (cross validation, bootstrapping, y-scrambling, test-splitting) External validation technique Availability of information for training/test set Dataset of chemicals for training/test set Descriptor values for training/test set Endpoint values for training/test set
5.	Mechanistic interpretation (where feasible)	Relevance of descriptors/structural features to mode of action, e.g., electrophiles for skin sensitization; LogKow modeling hydrophobicity for baseline narcosis in aquatic fish toxicity

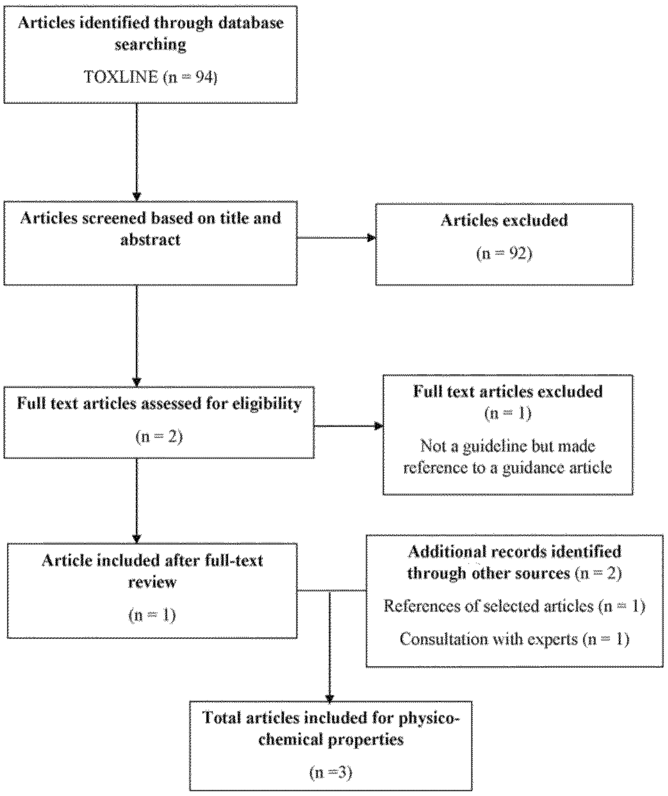

Articles identified through database searching
TOXLINE (n = 94)
Articles screened based on title and abstract
Articles excluded
(n = 92)
Full text articles assessed for eligibility
(n = 2)
Full text articles excluded
(n = 1)
Not a guideline but made reference to a guidance article
Article included after full-text review
(n = 1)
Additional records identified through other sources (n = 2)
References of selected articles (n = 1)
Consultation with experts (n = 1)
Total articles included for physico-chemical properties
(n =3)

Fig. 3. Flow of included guidance documents on physico-chemical studies.

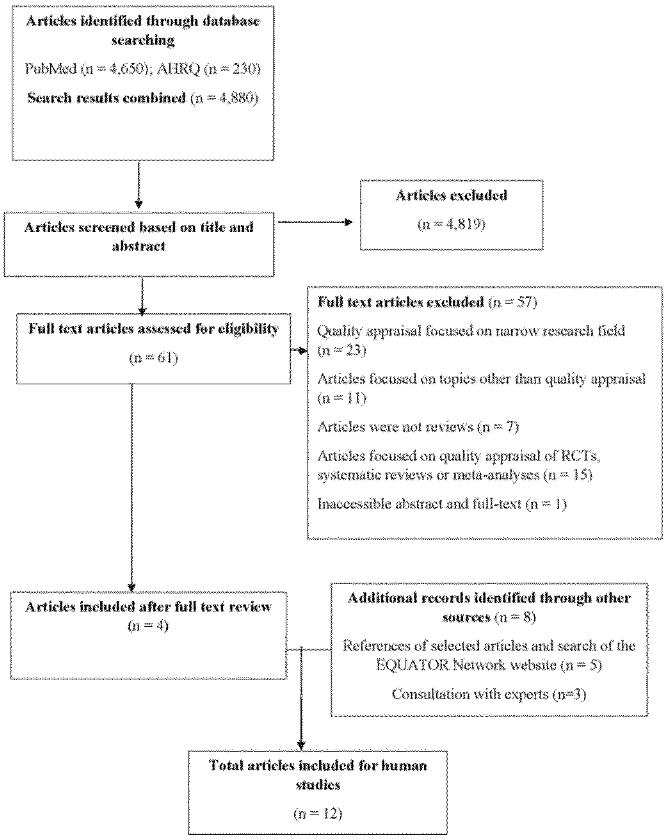

Articles identified through database searching
PubMed (n = 4,650); AHRQ (n = 230)
Search results combined (n = 4,880)
Articles excluded
(n = 4,819)
Articles screened based on title and abstract
Full text articles excluded (n = 57)
Quality appraisal focused on narrow research field (n = 23)
Articles focused on topics other than quality appraisal (n = 11)
Articles were not reviews (n = 7)
Articles focused on quality appraisal of RCTs, systematic reviews or meta-analyses (n = 15)
Inaccessible abstract and full-text (n = 1)
Full text articles assessed for eligibility
(n = 61)
Articles included after full text review
(n = 4)
Additional records identified through other sources (n = 8)
References of selected articles and search of the EQUATOR Network website (n = 5)
Consultation with experts (n=3)
Total articles included for human studies
(n = 12)

Fig. 4. Flow of included guidance documents on human observational studies.

3.4. Human studies

3.4.1. Literature search results

The literature search for guidance on the methodological and reporting quality of human studies returned 4880 citations. Preliminary screening of titles and abstracts yielded 61 papers for full-text review. Fifty-seven of these publications were excluded for a number of reasons (see Fig. 4). The following four documents met the eligibility criteria: Katrak et al. (2004), Mallen et al. (2006), Sanderson et al. (2007), and Viswanathan et al. (2008).

Five additional papers were retrieved from examining the references of the four selected publications, as well as from the EQUATOR Network's website: National Institute for Health and Care Excellence (2012), Harbour and Forsyth (2008), von Elm et al. (2007), Wells et al. (2004), and West et al. (2002). Three papers were added based on input from experts: Sterne et al. (2014), Money et al. (2013), and Lavelle et al. (2012). Consequently, 12 publications in total were identified as providing guidance on the assessment of human studies, including 10 on methodological quality, one on reporting quality, and one on mixed guidance. In addition, Rooney et al. (2014) and NRC (2014) (summarized above), which although not reviews of the subject, apply to human studies as well as in vivo and in vitro studies.

Throughout the literature search, we focused on observational studies, the type of human study most likely to be encountered in the context of hazard identification and risk assessment of environmental chemicals.

3.4.2. Methodological quality

Two papers that address the methodological quality of human studies in the context of risk assessment are summarized first, before moving on to eight papers on the same topic in the context of healthcare and biomedicine.

1. Lavelle et al. (2012)

This paper describes a framework for assessing the quality of human and animal data, in the context of integrating the two types of data into risk assessments. The framework is intended to allow risk assessors to evaluate the intrinsic strengths and weaknesses of each type of study and thereby select the most appropriate data for the risk assessment process. Seven criteria are proposed for assessing the quality of human studies (study design/conduct, subject selection, sample size and power, exposure assessment, outcome data, bias and confounding, and statistical analysis). Once assessed according to these criteria, studies are assigned to one of four quality categories, ranging from "A," if all seven methodological elements have been thoroughly addressed, to "D," if the study fails to meet the most basic standards important to epidemiologic research (e.g. an inappropriate study design for the research question). The resulting ratings of human data are compared to analogous ratings of available animal data in order to determine the most appropriate data for risk assessment.

2. Money et al. (2013)

Money et al. describe an approach to assess the methodological quality of human data in the context of risk assessments under REACH. Much of the available human data in this context are observational rather than experimental in nature. Money et al. divide the quality of human evidence into four categories, analogous to those for animal data quality proposed by Klimisch et al. (1997): Type 1 "(reliable without restriction "), Type 2 ("reliable with restriction"), Type 3 ("not reliable"), and Type 4 ("not assignable"). The categorizations are based on quality criteria that vary depending on whether the outcomes of interest are chronic or non-specific versus acute or specific. The consistency between the framework proposed for human studies and the original Klimisch framework for animal studies facilitates comparisons between the human and animal data. However, given the preference for human data, a human study with a quality category poorer than that of an animal in vivo study is not necessarily given less weight.

3. West et al. (2002)

This paper reviews scales and checklists intended to rate the strength of evidence for healthcare practices. Scales and checklists evaluating observational studies were regarded as high quality if they considered the following nine major domains: a focused study question, a description of the study population, the comparability of subjects, a clear definition of the exposure or intervention, clearly stated primary/secondary outcomes, an appropriate statistical analysis, measures of effect and precision used appropriately, stated conclusions supported by results, and reporting of funding/sponsorship. West et al. found many tools to be deficient in empirical documentation of the framework used in guidance development.

4. Wells et al. (2004)

Wells et al. devised the Newcastle–Ottawa Scale (NOS) for the quality appraisal of human observational studies. The NOS focuses on the design, content, and integration of quality assessments in the interpretation of findings from independent, non-randomized studies. Eight quality criteria were grouped into three categories: (1) the selection of study groups, (2) the comparability of the groups, and (3) the ascertainment of either the exposure (for case control studies) or the outcome (for cohort studies) of interest. A study is awarded a maximum of one star for each criterion within the selection and exposure groups, and a maximum of two stars is awarded for comparability. For cohort studies, for example, four criteria are evaluated in the selection category (the representativeness of the exposed cohort, the selection of the non-exposed cohort, ascertainment of exposure, and the demonstration that the outcome of interest was not present at the start of the study), one in the comparability category (the comparability of cohorts on the basis of the design or analysis), and three in the outcome category (the assessment of the outcome, the adequacy of follow-up duration, and the adequacy of the follow-up of cohorts).

5. Katrak et al. (2004)

This systematic review was conducted to evaluate the content, intent, construction, and psychometric properties of methodological appraisal tools for all types of human study designs. One hundred and twenty-eight tools were identified, 19 of which focused on observational studies. The items in the appraisal tools were grouped into one of 12 categories: study aims and justification, methodology used, sample selection, method of randomization, blinding, attrition, outcome measure characteristics, intervention details, method of data analyses, potential sources of bias, issues of external validity, and miscellaneous. The most frequently addressed items for observational studies fell in the categories of sample selection (e.g. comparability of participants at baseline) and data analyses (e.g. appropriate statistical analyses and sample size calculations). Ten of the 19 appraisal tools summed up the results of quality appraisal in a single numeric score by either an equal weighting system (where one point was allocated to each item fulfilled) or a weighted system (where fulfilled items were assigned various points depending on perceived relevance). The remaining appraisal tools did not involve a summary score but left the overall quality appraisal to the discretion of the “research consumer.” None of the tools to appraise observational research documented evidence of their validity and reliability.

6. Mallen et al. (2006)

This study examined the nature and extent of methodological quality assessment in systematic reviews of observational studies. The methodological quality criteria commonly assessed in these systematic reviews included the use of accurate and appropriate outcome measures, adjustment of confounding, the appropriate selection of controls, assessment of loss to follow-up, and appropriate statistical analysis. The authors concluded that no consensus exists for which quality assessment tool or specific criteria should be applied when evaluating observational studies.

7. Sanderson et al. (2007)

Sanderson et al. reviewed tools for assessing methodological quality and susceptibility to bias in human observational studies. Eighty-six tools were identified, including checklists and scales. These tools were assessed according to whether they addressed what the authors believed were “key” domains of bias. The selection of six key domains was influenced by the STROBE guidelines for reporting observational studies (see below) and included: appropriateness of methods for selection of study participants, appropriateness of methods for measuring exposure and outcome variables, appropriateness of design-specific sources of bias (excluding confounding), appropriateness of methods to control confounding, appropriateness of statistical methods (primary analysis of effects without confounding), and conflict of interest. Most of the evaluated tools (78–92%) addressed each of these domains, but only a few tools (4%) assessed conflict of interest.

8. Harbour and Forsyth (2008)

The Scottish Intercollegiate Guideline Network (SIGN) published this handbook to provide a framework for evaluating the validity of studies on healthcare improvement. This evaluation of strength of evidence was intended to inform grading of recommendations for clinical guideline development. The authors examined available guidance on quality appraisal of different types of human studies, and then worked with a larger group of scientists to develop methodology checklists for quality appraisal. The checklists developed for observational studies (cohort studies and case–control studies) consist of two sections: one for the internal validity appraisal and the second for an overall assessment of the study. Items assessing internal validity are grouped into four categories: selection of subjects (e.g. comparability of study groups with source population), assessment (e.g. standard and valid measurement of exposure status), confounding (e.g. identification of potential confounders in the design and analysis), and statistical analysis (e.g. provision of confidence intervals). Various levels of the strength of evidence can then be assigned to a study based on the study type and the number of quality criteria met.

9. Viswanathan et al. (2008)

The AHRQ provides guidance for assessing the risk of bias of individual human studies in the context of comparative effectiveness research in health care. Several common study designs were examined, including various types of observational studies: cohort, case-control, case series, and cross-sectional studies. Sixteen criteria were developed for assessing the risk of bias of the different types of studies, of which 14 were felt to be applicable to observational studies. These criteria were grouped under five potential sources of bias: selection bias (5 items), performance bias (2 items), attrition bias (1 item), detection bias (5 items), and reporting bias (1 item). Examples of the criteria applicable to human observational studies and the types of bias they address include: application of uniform inclusion and exclusion criteria to all comparison groups (selection bias); adherence to the intervention protocol (performance bias); appropriate handling of missing data due to attrition issues, e.g. loss to follow-up (attrition bias); assessment of exposure using valid and reliable measurement methods (detection bias); and ensuring that potential outcomes are pre-specified and reported (reporting bias).

10. Sterne et al. (2014)

This tool – A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI) – has been recently designed by the Cochrane Collaboration and intended for the appraisal of observational studies. The tool uses the term “intervention” to refer to “treatment” or “exposure” and is thus potentially applicable to toxicology studies. ACROBAT-NRSI comprises seven domains in which bias may be introduced, and these apply to one of the three levels of the design of non-randomized studies: pre-intervention, intervention, and post-intervention. Pre-intervention biases that are evaluated include bias due to confounding (selection bias) and bias in the selection of participants into the study (selection bias). At-intervention biases include bias in the measurement of intervention (observer bias, recall bias, etc.). Post-intervention biases include bias due to departures from intended interventions (performance bias), bias due to missing data (attrition bias), bias in the measurement of outcomes (detection bias), and bias in the selection of the reported result (outcome reporting bias). The response options for each of the seven domain-level judgments are “low,” “moderate,” “serious,” or “critical” risk of bias, as well as “no information” (documentation not available upon which to base a judgment).

3.4.3. Reporting quality

One study addressing reporting quality was found by searching the EQUATOR Network's website.

1. von Elm et al. (2007)

The STROBE statement, drafted by a group of methodologists, researchers, and journal editors, provides recommendations for complete and accurate reporting of human observational studies (i.e. cohort, case-control, and cross-sectional studies). The checklist consists of 22 items that cover standard sections of a scientific paper: title and abstract (e.g. informative abstract of what was done and found), introduction (e.g., scientific background and rationale for investigation), methods (e.g. key elements of study design), results (e.g. characteristics of study participants), discussion (e.g. key results with reference to study objectives), and other information (e.g. source of funding). Multiple extensions of the STROBE statement have now been developed for specific fields of study, including molecular epidemiology, genetic association, and infectious diseases.

3.4.4. Mixed guidance (methodological and reporting quality)

One paper was identified that provides a substantial mix of guidance on methodological and reporting quality of human observational studies.

1. The National Institute for Health and Care Excellence (2012)

The UK's National Institute for Health and Care Excellence (NICE) develops quality standards and performance metrics for providers of public health and social care services. Recommendations on quality appraisal of different types of studies were developed by a review of best available evidence in the literature, including the opinions of experts in healthcare. Methodology checklists for quality appraisal were developed for eight different types of studies, including observational studies. For observational (cohort) studies, the proposed checklist consists of items classified in four bias-assessment categories: selection bias (e.g. adequate allocation concealment), performance bias (e.g. blinding of participants and investigators to treatment allocation), attrition bias (e.g. follow-up of all study groups for an equal duration for differences), and detection bias (e.g. the use of a valid and reliable outcome measurement method). The methodology checklist for observational (case-control) studies is divided into two sections: criteria assessing internal validity and those assessing adequate reporting. Examples of criteria assessing internal validity include cases and controls from comparable populations and the control of potential confounders. Criteria assessing adequate reporting include the description of funding sources, the explanation of the size of effects observed, and the description of the main characteristics of the study population.

3.4.5. Criteria summary

Eleven guidance documents were identified that address the issue of the methodological quality of human observational studies (10 documents from the Methodological quality section and one from the Mixed guidance section). Most of these documents were from the clinical literature. There was substantial consistency across these documents in the proposed criteria for assessing methodological quality (Table 5). The most commonly proposed criteria were reliable exposure and outcome assessment, study design appraisal, comparability of group's baseline characteristics, statistical design evaluation, and loss to follow-up. Only one guidance document focused exclusively on assessing the reporting quality of human observational studies, which rendered an examination of trends impossible. This document, the authoritative STROBE statement, addresses criteria pertaining to the standard elements of a scientific paper.

4. Discussion and conclusions

This scoping review of guidance on methodological and reporting quality focuses on study types of direct relevance to toxicology and risk assessment, namely, in vivo, in vitro, (Q)SAR, physico-chemical, and human observational studies. Guidance on other study types, such as human controlled trials, may provide additional insights, but are beyond the scope of this review.

Table 5
Commonly proposed criteria for assessing the methodological quality of human observational studies from the “Methodological quality” and “Mixed guidance (methodological and reporting quality)” subsections of the “Human studies” section. These guidelines may assess a wider array of criteria; here, we list only those that have been commonly proposed.

Guidance	Reliable exposure & outcome measurement	Comparability of groups' baseline characteristics	Loss to follow-up	Statistical analysis	Sample size determination	Study design appraisal	Blinding of participants[a]
Sterne et al. (2014). (ACROBAT-NRSI Cochrane risk of bias assessment tool)	✓	✓	✓	✓	–	–	✓
Money et al. (2013). (Evaluating and scoring human data)	✓	✓	✓	✓	✓	✓	–
Lavelle et al. (2012). (Integrating human and animal data)	✓	✓	–	✓	✓	✓	–
National Institute for Health and Care Excellence (2012). (Methodology checklist)	✓	✓	✓	✓	–	–	✓
Harbour and Forsyth (2008). (SIGN 50 guideline developer's handbook)	✓	✓	✓	✓	✓	✓	✓
Viswanathan et al. (2008). (AHRQ, Assessing the risk of bias)	✓	✓	✓	–	–	–	–
Sanderson et al. (2007). (Assessing quality and susceptibility to bias)	✓	✓	✓	✓	–	✓	✓
Mallen et al. (2006). (Quality assessment of observational studies)	✓	✓	✓	✓	✓	✓	✓
Katrak et al. (2004). (The content of critical appraisal tools)	✓	✓	–	✓	✓	✓	✓
Wells et al. (2004). (Newcastle–Ottawa Scale)	✓	✓	✓	–	–	–	–
West et al. (2002). (Systems to Rate the Strength of Scientific Evidence)	✓	✓	✓	✓	✓	–	–
% of total (N = 11)	100	100	82	82	55	55	45

[a] In cohort studies, blinding ensures participants receiving care or individual administering care are kept “blind” to treatment allocation. In cases where blinding is not feasible, there is awareness that knowledge of exposure status could have affected outcome assessment.

Although grouping approaches, including read-across, are not specifically covered in this review, it is worth noting that the principles underpinning (Q)SAR studies are complementary to those for analogue and category approaches. Reporting formats akin to the QMRF and QPRF exist within the OECD QSAR Toolbox for the various techniques for filling data gaps. The OECD QSAR Toolbox is a software tool for the development, justification, and documentation of chemical categories. A separate reporting format to document the overall justification for the analogue/category approach is captured in the Analogue/Category Reporting Formats, which are described in the OECD grouping guidance (OECD, 2014; Patlewicz et al., 2014; ECHA, 2008).

The criteria most commonly proposed for assessing the methodological and reporting quality of toxicologically relevant studies are summarized in tabular form, where appropriate (Tables 2–5). In cases in which several guidance documents address a given study type, there is considerable overlap in the proposed criteria, despite some difference across guidance documents (Tables 2, 3, and 5). This is reassuring, as quality appraisals should ideally be based on consensus criteria in order to facilitate broad understanding, buy-in, and comparison across assessments, as well as to facilitate the conduct of the appraisals themselves. The results also illustrate that the proposed criteria differ somewhat across study types, suggesting that appraisal tools may need to be tailored to particular study types.

It is clear from this review that the available guidance on methodological and reporting quality is more extensive for in vivo and human observational studies than for in vitro, (Q)SAR, and physico-chemical studies. However, the available guidance for (Q)SARs and physico-chemical studies addresses the critically important consideration of regulatory compliance. Moreover, by extension, this guidance also has relevance to non-regulatory applications.

The claim that studies with high risk of bias can yield distorted outcomes has been demonstrated primarily in human medicine and healthcare (Nieto et al., 2007; Schulz et al., 1995). A recent overview of systematic reviews of animal efficacy studies across a wide range of outcomes and disease areas “demonstrates the need for randomization, allocation concealment, and blind outcome assessment in animal research” (Hirst et al., 2014). This conclusion underscores the importance of examining the methodological quality of animal studies. Similar investigations should be made in toxicology, to determine which quality criteria have the largest impacts on study outcomes in this field.

Similarly, seriously incomplete reporting can impede understanding of the research and contribute to research waste through unnecessary replication of poorly reported studies (Ioannidis et al., 2014). Poor reporting can also undermine assessments of methodological quality. In fact, we would argue that any criterion important enough to be proposed for methodological quality should also be regarded as a reporting criteria, further underscoring the linkage between methodological quality and reporting completeness.

Frameworks for appraising methodological quality are still emerging and in flux in toxicology, especially for in vivo and in vitro studies (Krauth et al., 2013; Rooney et al., 2014). Much is being borrowed from clinical medicine and healthcare, with their emphasis on risks of bias. We used the risk of bias framework for categorizing the proposed criteria for in vivo and in vitro studies, where it applied to roughly half of the criteria (Table 2). On the other hand, the applicability of this framework to quality assessment for (Q)SAR and physico-chemical studies has apparently yet to be explored.

Apart from a relatively new interest in the risk of bias framework, toxicology – especially regulatory toxicology – has had a longstanding emphasis on quality assurance (QA) and quality control (QC), including recommendations on personnel training, equipment calibration, record-keeping, chemical characterization, and the housing and feeding of animals. Toxicology also has a longstanding emphasis on protocol standardization through harmonized test guidelines and GLP standards. This historical emphasis on quality assurance, quality control, and standardization within toxicology should be integrated with the emerging emphasis on risk of bias into a coherent framework (Beck et al., 2014) and terminology should be harmonized across disciplines, where appropriate.

It should be borne in mind that the present work is an exploratory mapping of a rapidly evolving field. A number of methodological decisions were made to grapple with the challenges of this literature, such as the non-standardized terminology and criteria and the resulting ambiguities (see Methods section). Some level of subjectivity probably is unavoidable in any identification and categorization of criteria in this literature. In some cases, the criteria and classifications in our tables may be arguable. We nonetheless believe that the resulting mapping fulfills our primary goal of providing an entry point into this field.

The results should be taken as a starting point for further refinement. Ideally, guidance should be based not on how frequently a given criterion has been proposed, but on assessments of which criteria make the biggest contributions to outcomes for the study types and fields of interest. What are the high-impact criteria for methodological quality in toxicology? And for reporting quality/completeness? Once identified, these criteria should then be emphasized in subsequent iterations of guidance.

We have framed this review primarily in the context of applying guidance to the assessment of existing studies. Those interested in assessing the design, conduct, analysis, or reporting of existing studies include risk assessors and journal editors and reviewers, as well as scientists interested in understanding and possibly replicating a study or appraising the work of a given researcher (e.g. tenure review committees). Such assessments can also be conducted on planned studies by, for example, grant reviewers, funding agencies, and human and animal research review boards. Apart from assessments of existing or planned studies, researchers themselves can apply guidance on methodological and reporting quality to ensure the proper design, conduct, analysis, and reporting of their studies.

Three other relevant contexts are worthy of note. First, appraisals of methodological quality or risk of bias figure prominently in systematic reviews. Such appraisals are made of the individual studies included in a review and the results of these appraisals contribute to the overall “grade” of the “quality of evidence” in the review (Higgins and Green, 2008; Guyatt et al., 2011b; Rooney et al., 2014). Consequently, the approaches and principles discussed in this paper can help address the increasing calls for systematic reviews in literature-based chemical assessments (National Research Council, 2011; Thayer et al., 2012; Birnbaum et al., 2013), as well as broader calls for developing an evidence-based toxicology (Guzelian et al., 2005; Hoffmann and Hartung, 2005; Stephens et al., 2013). Second, greater attention to quality criteria can aid calls for enhancing reproducibility in animal studies (Collins & Tabak, 2014; Thayer et al., 2014). And third, regulatory programs have been important drivers of study appraisals in toxicology. Government programs that manage risks associated with new and existing substances, such as the REACH program in the EU and the High Production Volume and the IRIS programs of the Environmental Protection Agency (EPA) in the US, have called for assessments of the methodological quality of studies to be included in submitted dossiers (National Research Council, 2014; Christensen et al., 2011; Foth and Hayes, 2008; Tunkel et al., 2005; Green et al., 2001). Such mandates have led to a stronger emphasis on the use of existing guidance and indeed to the development of new guidance, summarized herein.

In sum, the guidance summarized herein has applicability to many aspects of toxicology. Greater attention to the methodological and reporting quality of toxicologically relevant studies has the potential to improve the science of toxicology and the resulting decision-making based on this science, as well as encourage more efficient spending on research and usage of animals in experimentation.

Acknowledgements

Authors GOS, SH, TH, and MLS wish to thank an anonymous philanthropic donor and the ExxonMobil Foundation for funding to the Johns

Hopkins Center for Alternatives to Animal Testing (CAAT) to carry out this project of the Evidence-based Toxicology Collaboration (http://www.ebtox.com/), for which CAAT serves as secretariat.

Appendix A. Methodological details regarding the literature searches

A.1. In vitro and in vivo studies

In PubMed and Embase, terms to capture in vitro and in vivo studies in toxicology or biomedical fields, and terms to capture guidance, were developed and applied.

A.1.1. PubMed

Terms to capture in vitro and in vivo studies in toxicology or biomedical fields, and terms to capture guidance, were combined by using "AND" as a Boolean string.

A.1.1.1. Terms to capture in vivo and in vitro studies in toxicology or biomedical fields. ("Models, Animal"[Majr] OR "Animal Experimentation"[Majr] OR "Animal research"[tw] OR "Animal Studies"[tw] OR "Animal Testing Alternatives/methods"[Majr] OR "Animals, Laboratory"[Majr] OR "Preclinical"[tw] OR "Toxicology/methods"[Majr] OR "Ecotoxicology/methods"[Majr] OR "Toxicology/classification"[Majr] OR "Cell Culture Techniques/methods"[Majr]) OR "Review Literature as Topic"[Majr].

A.1.1.2. Terms to capture guidance. ("Research Design/standards"[Majr] OR "Reproducibility of Results"[Majr] OR "Risk Assessment"[Majr] OR "Guidelines as Topic"[Majr] OR "Quality Control"[Majr] OR "Data Collection/standards"[Majr] OR "risk of bias"[tw] OR "quality of reporting"[tw] OR "reporting quality"[tw] OR "reliability"[tw] OR "validity"[tw]).

A.1.2. Embase

For Embase, terms to capture in vitro and in vivo studies in toxicology or biomedical fields, and terms to capture guidance, were developed as shown below:

1. 'in vitro study':de,ab,ti
2. 'in vivo study':de,ab,ti
3. 'Animal studies':de,ab,ti
4. 'Toxicology'/exp
5. 'Ecotoxicology'/exp
6. 'Drug screening'/exp
7. 'culture technique'/exp
8. #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7
9. 'risk of bias' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
10. 'Reporting' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
11. 'Validity' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
12. 'Reliability' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
13. 'Adequacy' NEAR/5 (guideline* OR guidance* OR recommendation* OR standard* OR tool* OR checklist* OR criteria*)
14. Good practice* NEAR/5 (guideline* OR guidance* OR recommendation* OR standard* OR tool* OR checklist* OR criteria*)
15. #9 OR #10 OR #11 OR #12 OR #13 OR #14
16. #15 AND #8
17. #16 AND [English]/lim.

A.1.3. Eligibility criteria

Retrieved papers were included in this review if they were original studies, and characterized themselves as guidelines, checklists, tools, or instruments for assessing reliability, risk of bias, validity, adequacy, or quality of conduct or reporting of in vivo and/or in vitro studies.

A.2. (Q)SAR studies

In PubMed and Embase, terms to capture (Q)SAR studies in toxicology or biomedical fields, and terms to capture guidance, were developed and applied.

A.2.1. PubMed

Terms to capture (Q)SAR studies and terms to capture guidance were combined by using "AND" as a Boolean string.

A.2.1.1. Terms to capture (Q)SARs. "QSAR"[tw] OR "QSARs"[tw] OR "quantitative structure activity relationship"[tw] OR "quantitative structure activity relationships"[tw] OR "Quantitative Structure-Activity Relationship"[Mesh] OR "Computer Simulation"[Mesh] OR "in silico toxicology"[tiab] OR "in silico modeling"[tiab] OR "in silico study"[tiab] OR "computer simulation"[tw] OR ("Structure-Activity Relationship"[Mesh] AND ("1972/01/01"[PDAT]: "2000/12/31"[PDAT])).

A.2.1.2. Terms to capture guidance. "Research Design/standards"[Majr] OR "Reproducibility of Results"[Majr] OR "Risk Assessment"[Majr] OR "Guidelines as Topic"[Majr] OR "Quality Control"[Majr] OR "Data Collection/standards"[Majr] OR "risk of bias"[tw] OR "quality of reporting"[tw] OR "reporting quality"[tw] OR "reliability"[tw] OR "validity"[tw].

A.2.2. Embase

For Embase, terms to capture (Q)SAR studies and terms to capture guidance were developed as shown below:

1. QSAR*:de,ab,ti
2. In silico:de,ab,ti
3. Quantitative structure activity relationship*
4. 'Quantitative structure activity relation'/exp
5. 'Computer simulation'/exp
6. #1 OR #2 OR #3 OR #4 OR #5
7. Valid* NEAR/5 (guideline OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
8. 'adequacy' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
9. 'reliability' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
10. 'reporting' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
11. 'risk of bias' NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
12. good AND practice* NEAR/5 (guideline* OR guidance OR recommendation* OR standard* OR tool* OR checklist* OR criteria)
13. #7 OR #8 OR #9 OR #10 OR #11 OR #12
14. #13 AND #6.

A.2.3. Eligibility criteria

Retrieved papers were included in the review if they: (1) referred to (Q)SARs, in silico modeling, or a computer simulation technique, (2) were original studies (not reviews, commentaries, meeting proceedings, or duplicate publications), and (3) characterized themselves as guidelines for assessing the quality of reporting, validity, or reliability of computer-based methods.

A.3. Studies of physico-chemical properties

Guidance on assessing physico-chemical property studies was identified using TOXLINE. The following search terms were used: Guideline* AND Test method* AND (Physicochemical Phenomena OR Physical

properties OR Chemical properties OR physicochemical properties). Studies were included if they: (1) referred to physico-chemical properties; (2) were part of the primary literature (not reviews, commentaries or meeting proceedings); and (3) characterized their aim as providing guidelines for assessing the
reliability or validity of physico-chemical properties.

A.4. Human studies

In PubMed, an electronic search was designed to capture papers by using a combination of terms in three domains:

1. Terms to capture categories of human studies: “Peer Review, Research/standards”[Mesh] OR “Epidemiologic Research Design”[Mesh] OR “Evidence-Based Medicine/methods”[Mesh] OR “Epidemiologic Studies”[Mesh] OR “Research Design”[Mesh] OR “Epidemiology” [Mesh] OR “Case control study”[tiab] OR “Case control studies”[tiab] OR “Cohort study”[tiab] OR “Cohort studies”[tiab] OR “Cross-sectional study”[tiab] OR “Cross-sectional studies”[tiab] OR “Longitudinal study”[tiab] OR “Longitudinal studies”[tiab] OR “Observational study”[tiab] OR “Observational studies”[tiab]
2. Terms to capture reviews: “Review Literature as Topic”[Mesh] OR “Meta-Analysisas Topic”[Mesh] OR “Systematic review”[tiab]
3. Terms to capture quality appraisal: “Quality Control”[Mesh] OR “Quality Assurance, Health Care”[Mesh] OR “Guidelines as Topic”[Mesh] OR “Evidence-Based Medicine/standards”[Mesh] OR “Reproducibility of Results”[Mesh] OR “Quality Indicators, Health Care”[Mesh] OR “Publishing/standards”[Mesh].

In addition to the literature search in PubMed, separate searches were conducted in the AHRQ review repository and on the EQUATOR Network's website. On the AHRQ website (http://www.effectivehealthcare.ahrq.gov), we entered “quality of reporting” OR “risk of bias” in the search box. We hand-searched eligible guidelines on the EQUATOR Network's website (http://www.equator-network.org/).

A.4.1. Eligibility criteria

Eligible papers were those that were narrative reviews or systematic reviews of guidelines or instruments (checklists, scales or tools) used for assessing methodological or reporting quality, not an individual proposal of an instrument or application of an existing instrument. Publications were excluded if they focused on quality appraisal of meta-analyses or systematic reviews, not of individual studies.

References

Ågerstrand, M., Küster, A., Bachmann, J., Breitholtz, M., Ebert, I., Rechenberg, B., Rudén, C., 2011. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. Environ. Pollut. 159 (10), 2487–2492. http://dx.doi.org/10.1016/j.envpol.2011.06.023

Arksey, H., O'Malley, L., 2005. Scoping studies: towards a methodological framework. Int. J. Soc. Res. Methodol. 8 (1), 19–32. http://dx.doi.org/10.1080/1364557032000119616

Arts, J.H.E., Muijser, H., Jonker, D., van de Sandt, J.J.M., Bos, P.M.J., Feron, V.J., 2008. Inhalation toxicity studies: OECD guidelines in relation to REACH and scientific developments. Exp Toxicol. Pathol. 60 (2-3), 125–133. http://dx.doi.org/10.1016/j.etp.2008.01.011.

Bailoo, J.D., Reichlin, T.S., Würbel, H., 2014. Refinement of experimental design and conduct in laboratory animal research. ILAR J. Natl. Res. Counc. Inst. Lab. Anim. Resour. 55 (3), 383–391. http://dx.doi.org/10.1093/ilar/ilu037

Beck, N.B., Becker, R.A., Boobis, A., Fergusson, D., Fowle, J.R., Goodman, J., … Stephens, M.L., 2014. Instruments for assessing risk of bias and other methodological criteria of animal studies: omission of well-established methods. Environ. Health Perspect. 122 (3), A66–A67. http://dx.doi.org/10.1289/ehp.1307727

Beronius, A., Molander, L., Rudén, C., Hanberg, A., 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. J. Appl. Toxicol. 34 (6), 607–617. http://dx.doi.org/10.1002/jat.2991

Birnbaum, L.S., Thayer, K.A., Bucher, J.R., Wolfe, M.S., 2013. Implementing systematic review at the national toxicology program: status and next steps. Environ. Health Perspect. 121 (4), a108–a109. http://dx.doi.org/10.1289/ehp.1306711

Christensen, F.M., Eisenreich, S.J., Rasmussen, K., Sintes, J.R., Sokull-Kluettgen, B., Van de Plassche, E.J., 2011. European experience in chemicals management: integrating science into policy. Environ. Sci. Technol. 45 (1), 80–89. http://dx.doi.org/10.1021/es101541b

Code of Federal Regulations. Title 40 — Protection of Environment Part 160 and Part 792. (2011). https://www.gpo.gov/fdsys/pkg/CFR-2011-title40-vol24/xml/CFR-2011-title40-vol24-part160.xmlhttps://www.gpo.gov/fdsys/pkg/CFR-2011-title40-vol32/xml/CFR-2011-title40-vol32-part792.xml

Coecke, S., Balls, M., Bowe, G., Davis, J., Gstraunthaler, G., Hartung, T., … Stokes, W., 2005. Guidance on good cell culture practice. a report of the second ECVAM task force on good cell culture practice. Altern. Lab. Anim. 33 (3), 261–287.

Collins, F.S., Tabak, L.A., 2014. Policy: NIH plans to enhance reproducibility. Nature 505 (7485), 612–613.

Colquhoun, H.L., Levac, D., O'Brien, K.K., Straus, S., Tricco, A.C., Perrier, L., … Moher, D., 2014. Scoping reviews: time for clarity in definition, methods, and reporting. J. Clin. Epidemiol. 67 (12), 1291–1294. http://dx.doi.org/10.1016/j.jclinepi.2014.03.013

Durda, J.L., Preziosi, D.V., 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. Hum. Ecol. Risk. Assess. 6 (5), 747–765. http://dx.doi.org/10.1080/10807030091124176

ECHA, 2008. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R. 6: QSARs and Grouping of Chemicals. Retrieved from http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf

EPA (undated-a). United States Environmental Protection Agency. Good Laboratory Practices Standards Compliance Monitoring Program. Resources and Guidance documents. 40 CFR part 160 — FIFRA & 40 CFR part 792 — TSCA. Retrieved from http://www.epa.gov/compliance/good-laboratory-practices-standards-compliance-monitoring-program Access date: May 25, 2014.

EPA (undated-b) United States Environmental Protection Agency OPPTS Harmonized Test Guidelines, Series 830. Retrieved from http://www.regulations.gov/#!docketBrowser;rpp = 25;po = 0;D = EPA-HQ-OPPT-2009-0151 Access date: December 17, 2015.

European Union, 2008. Regulations. Council Regulation (EC) No 440/2008 of 30 May 2008. Part A: Methods for the Determination of Physico-chemical Properties (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri = OJ:L:2008:142:0001:0739:en:PDF

Festing, M.F.W., Altman, D.G., 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. Nat. Res. Counc. Inst. Lab. Anim. Resour. 43 (4), 244–258.

Foth, H., Hayes, A., 2008. Background of REACH in EU regulations on evaluation of chemicals. Hum. Exp. Toxicol. 27 (6), 443–461. http://dx.doi.org/10.1177/0960327108092296

Green, S., Goldberg, A.M., Zurlo, J., 2001. The TestSmart-HPV Program—development of an integrated approach for testing high production volume chemicals. Regul. Toxicol. Pharmacol. 33 (2), 105–109. http://dx.doi.org/10.1006/rtph.2000.1435

Grimes, D.A., Schulz, K.F., 2002. Bias and causal associations in observational research. Lancet (Lond., Engl.) 359 (9302), 248–252. http://dx.doi.org/10.1016/S0140-6736(02)07451-2

Guyatt, G.H., Oxman, A.D., Montori, V., Vist, G., Kunz, R., Brozek, J., … Schünemann, H.J., 2011a. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J. Clin. Epidemiol. 64 (12), 1277–1282. http://dx.doi.org/10.1016/j.jclinepi.2011.01.011

Guyatt, G., Oxman, A.D., Akl, E.A., Kunz, R., Vist, G., Brozek, J., … Schünemann, H.J., 2011b. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J. Clin. Epidemiol. 64 (4), 383–394. http://dx.doi.org/10.1016/j.jclinepi.2010.04.026

Guzelian, P.S., Victoroff, M.S., Halmes, N.C., James, R.C., Guzelian, C.P., 2005. Evidence-based toxicology: a comprehensive framework for causation. Hum. Exp. Toxicol. 24 (4), 161–201.

Harbour, R., Forsyth, L., 2008. SIGN 50: A Guideline Developer's Handbook. Retrieved May 20, 2014, from http://www.sign.ac.uk/guidelines/fulltext/50/

Henderson, V.C., Kimmelman, J., Fergusson, D., Grimshaw, J.M., Hackam, D.G., 2013. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 10 (7), e1001489. http://dx.doi.org/10.1371/journal.pmed.1001489

Higgins, J.P., Green, S. (Eds.), 2008. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series (Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9780470712184;jsessionid = 8D384B57DC5A9585C2E234DBC8621EC6.d04t02

Hirst, J.A., Howick, J., Aronson, J.K., Roberts, N., Perera, R., Koshiaris, C., Heneghan, C., 2014. The need for randomization in animal trials: an overview of systematic reviews. PLoS One 9 (6), e98856. http://dx.doi.org/10.1371/journal.pone.0098856

Hobbs, D.A., Warne, M.S.J., Markich, S.J., 2005. Evaluation of criteria used to assess the quality of aquatic toxicity data. Integr. Environ. Assess. Manag. 1 (3), 174–180.

Hoffmann, S., Hartung, T., 2005. Diagnosis: toxic!—trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. Toxicol. Sci. 85 (1), 422–428. http://dx.doi.org/10.1093/toxsci/kfi099.

Hooijmans, C.R.C., Leenaars, M.M., Ritskes-Hoitinga, M.M., 2010. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. Alternatives to Laboratory Animals: ATLA 38 (2), 167–182. http://dx.doi.org/10.1258/la.2010.009117

Hooijmans, C.R., Rovers, M.M., de Vries, R.B.M., Leenaars, M., Ritskes-Hoitinga, M., Langendam, M.W., 2014. SYRCLE's risk of bias tool for animal studies. BMC Med. Res. Methodol. 14, 43. http://dx.doi.org/10.1186/1471-2288-14-43

Hulzebos, E., Gunnarsdottir, S., Rila, J.-P., Dang, Z., Rorije, E., 2010. An Integrated Assessment Scheme for assessing the adequacy of (eco)toxicological data under REACH. Toxicol. Lett. 198 (2), 255–262. http://dx.doi.org/10.1016/j.toxlet.2010.07.004

Ioannidis, J.P.A., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., … Tibshirani, R., 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383 (9912), 166–175. http://dx.doi.org/10.1016/S0140-6736(13)62227-8.

Jaworska, J.S., Comber, M., Auer, C., Van Leeuwen, C.J., 2003. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environ. Health Perspect. 111 (10), 1358–1360.

Katrak, P., Bialocerkowski, A.E., Massy-Westropp, N., Kumar, S., Grimmer, K.A., 2004. A systematic review of the content of critical appraisal tools. BMC Med. Res. Methodol. 4, 22. http://dx.doi.org/10.1186/1471-2288-4-22

Kilkenny, C., Browne, W., Cuthill, I., Emerson, M., Altman, D., 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. J. Pharmacol. Pharmacother. 1 (2), 94. http://dx.doi.org/10.4103/0976-500X.72351

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25 (1), 1–5. http://dx.doi.org/10.1006/rtph.1996.1076

Krauth, D., Woodruff, T.J., Bero, L., 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ. Health Perspect. http://dx.doi.org/10.1289/ehp.1206389

Küster, A., Bachmann, J., Brandt, U., Ebert, I., Hickmann, S., Klein-Goedicke, J., ... Rechenberg, B., 2009. Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. Regul. Toxicol. Pharmacol. 55 (3), 276–280. http://dx.doi.org/10.1016/j.yrtph.2009.07.005

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., ... Silberberg, S.D., 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490 (7419), 187–191. http://dx.doi.org/10.1038/nature11556

Lavelle, K.S., Robert Schnatter, A., Travis, K.Z., Swaen, G.M.H., Pallapies, D., Money, C., ... Vrijhof, H., 2012. Framework for integrating human and animal data in chemical risk assessment. Regul. Toxicol. Pharmacol. 62 (2), 302–312. http://dx.doi.org/10.1016/j.yrtph.2011.10.009

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., ... Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. 62 (10), e1–34. http://dx.doi.org/10.1016/j.jclinepi.2009.06.006

Macleod, M.R., Fisher, M., O'Collins, V., Sena, E.S., Dirnagl, U., Bath, P.M.W., ... Howells, D.W., 2009. Good laboratory practice preventing introduction of bias at the bench. Stroke 40 (3), e50–e52. http://dx.doi.org/10.1161/STROKEAHA.108.525386

Mallen, C., Peat, G., Croft, P., 2006. Quality assessment of observational studies is not commonplace in systematic reviews. J. Clin. Epidemiol. 59 (8), 765–769. http://dx.doi.org/10.1016/j.jclinepi.2005.12.010

Maxim, L., van der Sluijs, J.P., 2014. Qualichem in vivo: a tool for assessing the quality of in vivo studies and its application for bisphenol a. PLoS One 9 (1), e87738. http://dx.doi.org/10.1371/journal.pone.0087738

Moher, D., 2015. Endorsing and implementing reporting guidelines in journals. Int. J. Nurs. Stud. 52 (8), 1404–1405. http://dx.doi.org/10.1016/j.ijnurstu.2015.04.015

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P., Walsh, S., 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control. Clin. Trials 16 (1), 62–73.

Money, C.D., Tomenson, J.A., Penman, M.G., Boogaard, P.J., Jeffrey Lewis, R., 2013. A systematic approach for evaluating and scoring human data. Regul. Toxicol. Pharmacol. 66 (2), 241–247. http://dx.doi.org/10.1016/j.yrtph.2013.03.011

National Research Council, 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. The National Academies Press Retrieved from http://www.nap.edu/openbook.php?record_id = 13142

National Research Council, 2014. Review of EPA's Integrated Risk Information System (IRIS) Process |. The National Academies Press (Retrieved July 31, 2015, from http://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process).

Nieto, A., Mazon, A., Pamies, R., Linana, J.J., Lanuza, A., Jiménez, F.O., ... Nieto, F.J., 2007. Adverse effects of inhaled corticosteroids in funded and nonfunded studies. Arch. Intern. Med. 167 (19), 2047–2053. http://dx.doi.org/10.1001/archinte.167.19.2047

NIH. National Institutes of Health (undated). Principles and Guidelines for Reporting Preclinical Research. Retrieved from http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research Access date: December 17, 2015.

O'Connor, A.M., Sargeant, J.M., 2014. Critical appraisal of studies using laboratory animal models. ILAR J. Nat. Res. Counc., Inst. Lab. Anim. Resour. 55 (3), 405–417. http://dx.doi.org/10.1093/ilar/ilu038

OECD, 1998. OECD Environmental Health and Safety Publications Series on Principles of Good Laboratory Practice and Compliance Monitoring No. 1 OECD Principles of Good Laboratory Practice (as Revised in 1997). Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote = env/mc/chem%2898%2917&doclanguage = en Access date: May 24, 2014.

OECD, 2007. Guidance document on the validation of quantitative structure–activity relationship [QSAR] models. OECD Environment Health and Safety Publication, Paris (2007) (Series on testing and assessment no. 69, ENV/JM/MONO(2007)2) Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote = env/jm/mono%282007%292&doclanguage = en Access date: May 25, 2014).

OECD, 2014. Guidance on Grouping of Chemicals. OECD Series on Testing and Assessment No. 194 Organisation for Economic Co-operation and Development, Paris, France (Retrieved January 2, 2015, from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote = env/jm/mono%282014%294&doclanguage = en

OECD (undated-a) OECD Guidelines for the Testing of Chemicals. Retrieved from http://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm Access date: May 24, 2014.

OECD (undated-b) OECD Guidelines for the Testing of Chemicals, Section 1. Physical–Chemical properties. Retrieved from http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-1-physical-chemical-properties_20745753 Access date: May 25, 2014

Olivo, S.A., Macedo, L.G., Gadotti, I.C., Fuentes, J., Stanton, T., Magee, D.J., 2008. Scales to assess the quality of randomized controlled trials: a systematic review. Phys. Ther. 88 (2), 156–175. http://dx.doi.org/10.2522/ptj.20070147

Patlewicz, G., Ball, N., Becker, R.A., Booth, E.D., Cronin, M.T.D., Kroese, D., ... Hartung, T., 2014. Read-across approaches—misconceptions, promises and challenges ahead. ALTEX 31 (4), 387–396.

Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R., Thayer, K.A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ. Health Perspect. 122 (7), 711–718. http://dx.doi.org/10.1289/ehp.1307972.

Sanderson, S., Tatt, I.D., Higgins, J.P.T., 2007. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int. J. Epidemiol. 36 (3), 666–676. http://dx.doi.org/10.1093/ije/dym018.

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., ... Hoffmann, S., 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicol. Lett. 189 (2), 138–144. http://dx.doi.org/10.1016/j.toxlet.2009.05.013

Schulz, K.F., Chalmers, I., Hayes, R.J., Altman, D.G., 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273 (5), 408–412.

Shea, B.J., Grimshaw, J.M., Wells, G.A., Boers, M., Andersson, N., Hamel, C., ... Bouter, L.M., 2007. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med. Res. Methodol. 7, 10. http://dx.doi.org/10.1186/1471-2288-7-10.

Stephens, M.L., Andersen, M., Becker, R.A., Betts, K., Boekelheide, K., Carney, E., ... Zurlo, J., 2013. Evidence-based toxicology for the 21st century: opportunities and challenges. ALTEX 30 (1), 74–104.

Sterne, J.A.C., Higgins, J.P.T., Reeves, B.C., on behalf of the development group for ACROBAT- NRSI, 2014. A Cochrane Risk of Bias Assessment Tool: for Non-randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0, 24 September 2014 Available from http://www.riskofbias.info (accessed 04/12/2015).

Thayer, K.A., Heindel, J.J., Bucher, J.R., Gallo, M.A., 2012. Role of environmental chemicals in diabetes and obesity: a National Toxicology Program workshop review. Environ. Health Perspect. 120 (6), 779–789. http://dx.doi.org/10.1289/ehp.1104597

Thayer, K.A., Wolfe, M.S., Rooney, A.A., Boyles, A.L., Bucher, J.R., Birnbaum, L.S., 2014. Intersection of systematic review methodology with the NIH reproducibility initiative. Environ. Health Perspect. 122 (7), A176–A177. http://dx.doi.org/10.1289/ehp.1408671.

The National Institute for Health and Care Excellence. (2012). The Guidelines Manual: Appendices B–I. http://publications.nice.org.uk/the-guidelines-manual-appendices-bi-pmg6b/appendix-d-methodology-checklist-cohort-studies http://publications.nice.org.uk/the-guidelines-manual-appendices-bi-pmg6b/appendix-e-methodology-checklist-casecontrol-studies

Tunkel, J., Mayo, K., Austin, C., Hickerson, A., Howard, P., 2005. Practical considerations on the use of predictive models for regulatory purposes. Environ. Sci. Technol. 39 (7), 2188–2199.

Unger, E.F., 2007. All is not well in the world of translational research. J. Am. Coll. Cardiol. 50 (8), 738–740. http://dx.doi.org/10.1016/j.jacc.2007.04.067

van der Worp, H.B., Howells, D.W., Sena, E.S., Porritt, M.J., Rewell, S., O'Collins, V., Macleod, M.R., 2010. Can animal models of disease reliably inform human studies? PLoS Med. 7 (3), e1000245. http://dx.doi.org/10.1371/journal.pmed.1000245

van Luijk, J., Bakker, B., Rovers, M.M., Ritskes-Hoitinga, M., de Vries, R.B.M., Leenaars, M., 2014. Systematic reviews of animal studies; missing link in translational research? PLoS One 9 (3), e89981. http://dx.doi.org/10.1371/journal.pone.0089981

Viswanathan, M., Ansari, M.T., Berkman, N.D., Chang, S., Hartling, L., McPheeters, M., ... Treadwell, J.R., 2008. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality (US), Rockville (MD) (Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK91433/

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., Vandenbroucke, J.P., Initiative, S.T.R.O.B.E., 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 335 (7624), 806–808. http://dx.doi.org/10.1136/bmj.39335.541782.AD

Wells, Shea, Connell, Peterson, Welch, Losos, & Tugwell. (2004). Ottawa Hospital Research Institute. Retrieved May 18, 2014, from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F., Lux, L., 2002, Marchh. Systems to Rate the Strength of Scientific Evidence: Summary [Text]. Retrieved March 1, 2014, from http://www.ncbi.nlm.nih.gov/books/NBK11930/

Worth, A., Bassan, A., Gallegos, A., Netzeva, T., Patlewicz, G., Pavan, M., ... Vracko, M., 2005. The Characterisation of (Q) SARs: Preliminary Guidance. JRC Report EUR 21866 EN European Chemicals Bureau, Joint Research Center, European Commission, Ispra, Italy.